# 1

# Observational Studies

## 1.1 What Are Observational Studies?

William G. Cochran first presented "observational studies" as a topic defined by principles and methods of statistics. Cochran had been an author of the 1964 United States Surgeon General's Advisory Committee Report, *Smoking and Health*, which reviewed a vast literature and concluded: "Cigarette smoking is causally related to lung cancer in men; the magnitude of the effect of cigarette smoking far outweighs all other factors. The data for women, though less extensive, point in the same direction (p. 37)." Though there had been some experiments confined to laboratory animals, the direct evidence linking smoking with human health came from observational or nonexperimental studies.

In a later review, Cochran (1965) defined an observational study as an empiric investigation in which:

> ... the objective is to elucidate cause-and-effect relationships ... [in which] it is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures.

Features of this definition deserve emphasis. An observational study concerns treatments, interventions, or policies and the effects they cause, and in this respect it resembles an experiment. A study without a treatment is neither an experiment nor an observational study. Most public opinion polls, most forecasting efforts, most studies of fairness and discrimination, and many other important empirical studies are neither experiments nor observational studies.

In an experiment, the assignment of treatments to subjects is controlled by the experimenter, who ensures that subjects receiving different treatments are comparable. In an observational study, this control is absent for one of several reasons. It may be that the treatment, perhaps cigarette smoking or radon gas, is harmful and cannot be given to human subjects for experimental purposes. Or the treatment may be controlled by a political process that, perhaps quite appropriately, will not yield control merely for an experiment, as is true of much of macroeconomic and fiscal policy. Or the treatment may be beyond the legal reach of experimental manipulation even by a government, as is true of many management decisions in a private economy. Or experimental subjects may have such strong attachments to particular treatments that they refuse to cede control to an experimenter, as is sometimes true in areas ranging from diet and exercise to bilingual education. In each case, the investigator does not control the assignment of treatments and cannot ensure that similar subjects receive different treatments.

## 1.2 Some Observational Studies

It is encouraging to recall cases, such as *Smoking and Health*, in which observational studies established important truths, but an understanding of the key issues in observational studies begins elsewhere. Observational data have often led competent honest scientists to false and harmful conclusions, as was the case with Vitamin C as a treatment for advanced cancer.

### *Vitamin C and Treatment of Advanced Cancer: An Observational Study and an Experiment Compared*

In 1976, in their article in the *Proceedings of the National Academy of Sciences*, Cameron and Pauling presented observational data concerning the use of vitamin C as a treatment for advanced cancer. They gave vitamin C to 100 patients believed to be terminally ill from advanced cancer and studied subsequent survival.

For each such patient, 10 historical controls were selected of the same age and gender, the same site of primary cancer, and the same histological tumor type. This method of selecting controls is called *matched sampling*—it consists of choosing controls one at a time to be similar to individual treated subjects in terms of characteristics measured prior to treatment. Used effectively, matched sampling often creates treated and control groups that are comparable in terms of the variables used in matching, though the groups may still differ in other ways, including ways that were not measured. Cameron and Pauling (1976, p. 3685) write: "Even though no formal process of randomization was carried out in the selection of our two groups, we believe that they come close to representing random subpopulations of the population of terminal cancer patients in the Vale of Leven Hospital." In a moment, we shall see whether this is so.

Patients receiving vitamin C were compared to controls in terms of time from "untreatability by standard therapies" to death. Cameron and Pauling found that, as a group, patients receiving vitamin C survived about four times longer than the controls. The difference was highly significant in a conventional statistical test, $p$-value $< 0.0001$, and so could not be attributed to "chance." Cameron and Pauling "conclude that there is strong evidence that treatment ... [with vitamin C] ... increases the survival time."

This study created interest in vitamin C as a treatment. In response, the Mayo Clinic (Moertel et al., 1985) conducted a careful randomized controlled experiment comparing vitamin C to placebo for patients with advanced cancer of the colon and rectum. In a *randomized experiment*, subjects are assigned to treatment or control on the basis of a chance mechanism, typically a random number generator, so it is only luck that determines who receives the treatment. They found no indication that vitamin C prolonged survival, with the placebo group surviving slightly but not significantly longer. Today, few scientists claim that vitamin C holds promise as a treatment for cancer.

What went wrong in Cameron and Pauling's observational study? Why were their findings so different from those of the randomized experiment? Could their mistake have been avoided in any way other than by conducting a true experiment?

Definite answers are not known, and in all likelihood will never be known. Evidently, the controls used in their observational study, though matched on several important variables, nonetheless differed from treated patients in some way that was important to survival.

The obvious difference between the experiment and the observational study was the random assignment of treatments. In the experiment, a single group of patients was divided into a treated and a control group using a random device. Bad luck could, in principle, make the treated and control groups differ in important ways, but it is not difficult to quantify the potential impact of bad luck and to distinguish it from an effect of the treatment. Common statistical tests and confidence intervals do precisely this. In fact, this is what it means to say that the difference could not reasonably be due to "chance." Chapter 2 discusses the link between statistical inference and random assignment of treatments.

In the observational study, subjects were not assigned to treatment or control by a random device created by an experimenter. The matched sampling ensured that the two groups were comparable in a few important ways, but beyond this, there was little to ensure comparability. If the groups were not comparable before treatment, if they differed in important ways, then the difference in survival might be no more than a reflection of these initial differences.

It is worse than this. In the observational study, the control group was formed from records of patients already dead, while the treated patients were alive at the start of the study. The argument was that the treated patients were terminally ill, that they would all be dead shortly, so the recent records of apparently similar patients, now dead, could reasonably be used to indicate the duration of survival absent treatment with vitamin C. Nonetheless, when the results were analyzed, some patients given vitamin C were still alive; that is, their survival times were censored. This might reflect dramatic effects of vitamin C, but it might instead reflect some imprecision in judgments about who is terminally ill and how long a patient is likely to survive, that is, imprecision about the initial prognosis of patients in the treated group. In contrast, in the control group, one can say with total confidence, without reservation or caveat, that the prognosis of a patient already dead is not good. In the experiment, all patients in both treated and control groups were initially alive.

It is worse still. While death is a relatively unambiguous event, the time from "untreatability by standard therapies" to death depends also on the time of "untreatability." In the observational study, treated patients were judged, at the start of treatment with vitamin C, to be untreatable by other therapies. For controls, a date of untreatability was determined from records. It is possible that these two different processes would produce the same number, but it is by no means certain. In contrast, in the experiment, the starting date in treated and control groups was defined in the same way for both groups, simply because the starting date was determined before a subject was assigned to treatment or control.

What do we conclude from the studies of vitamin C? First, observational studies and experiments can yield very different conclusions. When this happens, the experiments tend to be believed. Chapter 2 develops some of the reasons why this tendency is reasonable. Second, matching and similar adjustments in observational studies, though often useful, do not ensure that treated and control groups are comparable in all relevant ways. More than this, the groups may not be comparable and yet the data we have may fail to reveal this. This issue is discussed extensively in later chapters. Third, while a controlled experiment uses randomization and an observational study does not, experimental control also helps in other ways. Even if we cannot randomize, we wish to exert as much experimental control as is possible, for instance, using the same eligibility criteria for treated and control groups, and the same methods for determining measurements.

Observational studies are typically conducted when experimentation is not possible. Direct comparisons of experiments and observational studies are less common, vitamin C for cancer being an exception. Another direct comparison occurred in the Salk vaccine for polio, a story that is well told by Meier (1972). Others are discussed by Chalmers, Block, and Lee (1970), LaLonde (1986), Fraker and Maynard (1987), Zwick (1991), Friedlander and Robins (1995), and Dehejia and Wahba (1999).

### *Smoking and Heart Disease: An Elaborate Theory*

Doll and Hill (1966) studied the mortality from heart disease of British doctors with various smoking behaviors. While dramatic associations are typically found between smoking and lung cancer, much weaker associations are found with heart disease. Still, since heart disease is a far more common cause of death, even modest increases in risk involve large numbers of deaths.

The first thing Doll and Hill did was to "adjust for age." The old are at greater risk of heart disease than the young. As a group, the smokers tended to be somewhat older than the nonsmokers, though of course there were many young smokers and many old nonsmokers. Compare smokers and nonsmokers directly, ignoring age, and you compare a somewhat older group to a somewhat younger group, so you expect a difference in coronary mortality even if smoking has no effect. In its essence, to "adjust for age" is to compare smokers and nonsmokers of the same age. Often results at different ages are combined into a single number called an age-adjusted mortality rate. Methods of adjustment and their properties are discussed in Chapters 3 and 10. For now, it suffices to say that differences in Doll and Hill's age-adjusted mortality rates cannot be attributed to differences in age, for they were formed by comparing smokers and nonsmokers of the same age. Adjustments of this sort, for age or other variables, are central to the analysis of observational data.

The second thing Doll and Hill did was to consider in detail what should be seen if, in fact, smoking causes coronary disease. Certainly, increased deaths among smokers are expected, but it is possible to be more specific. Light smokers should have mortality somewhere between that of nonsmokers and heavy smokers. People who quit smoking should also have risks between those of nonsmokers and heavy smokers, though it is not clear what to expect when comparing continuing light smokers to people who quit heavy smoking.

Why be specific? Why spell out in advance what a treatment effect should look like? The importance of highly specific theories has a long history, having been advocated in general by Sir Karl Popper (1959) and in observational studies by Sir Ronald Fisher, the inventor of randomized experiments, as quoted by Cochran (1965, §5):

> About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: 'Make your theories

TABLE 1.1. Coronary Mortality in Relation to Smoking.

$$\begin{tikzcd}
 & \substack{\text{Heavy Smokers} \\ 3.79} & \\
\substack{\text{Moderate Smokers} \\ 2.81} \arrow[ur] & & \substack{\text{Exsmokers} \\ 2.76} \arrow[ul] \\
\substack{\text{Light Smokers} \\ 2.72} \arrow[u] & & \\
 & \substack{\text{Nonsmokers} \\ 2.12} \arrow[ul] \arrow[uur] &
\end{tikzcd}$$

> elaborate.' The reply puzzled me at first, since by Occam's razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold.
>
> ... this multi-phasic attack is one of the most potent weapons in observational studies.

Chapters 6 through 9 consider this advice formally and in detail.

Table 1.1 gives Doll and Hill's six age-adjusted mortality rates for death from coronary disease not associated with any other specific disease. The rates are deaths per 1000 per year, so the value 3.79 means about 4 deaths in each 1000 doctors each year. The six groups are nonsmokers, exsmokers, and light smokers of 1 to 14 cigarettes, moderate smokers of 15 to 24 cigarettes, and heavy smokers of 25 or more cigarettes per day. Doll and Hill did not separate exsmokers by the amount they had previously smoked, though this would have been interesting and would have permitted more detailed predictions. Again, differences in age do not affect these mortality rates.

Table 1.1 confirms each expectation. Mortality increases with the quantity smoked. Quitters have lower mortality than heavy smokers but higher mortality than nonsmokers. Any alternative explanation, any claim that smoking is not a cause of coronary mortality, would need to explain the entire pattern in Table 1.1. Alternative explanations are not difficult to imagine, but the pattern in Table 1.1 restricts their number.

### *DES and Vaginal Cancer: Sensitivity to Bias*

Cancer of the vagina is a rare condition, particularly in young women. In 1971, Herbst, Ulfelder, and Poskanzer published a report describing eight cases of vaginal cancer in women aged 15 to 22. They were particularly interested in the possibility that a drug, diethylstilbestrol or DES, given to pregnant women, might be a cause of vaginal cancer in their daughters. Each of the eight cases was matched to four *referents*, that is, to four women who did not develop vaginal cancer. These four referents were born within five days of the birth of the case at the same hospital, and on the same type of service, ward or private. There were then eight cases of vaginal cancer and 32 referents, and the study compared the use of DES by their mothers.

This sort of study is called a *case-referent study* or a *case-control study* or a *retrospective study*, no one terminology being universally accepted. In an experiment and in many observational studies, treated and control groups are followed forward in time to see how outcomes develop. In the current context, this would mean comparing two groups of women, a treated group whose mothers had received DES and a control group whose mothers had not. That sort of study is not practical because the outcome, vaginal cancer, is so rare—the treated and control groups would have to be enormous and continue for many years to yield eight cases of vaginal cancer. In a case-referent study, the groups compared are not defined by whether or not they received the treatment, but rather by whether or not they exhibit the outcome. The cases are compared to the referents to see if exposure to the treatment is more common among cases.

In general, the name "case-control" study is not ideal because the word "control" does not have its usual meaning of a person who did not receive the treatment. In fact, in most case-referent studies, many referents did receive the treatment. The name "retrospective" study is not ideal because there are observational studies in which data on entire treated and control groups are collected after treatments have been given and outcomes have appeared, that is, collected retrospectively, and yet the groups being compared are still treated and untreated groups. See MacMahon and Pugh (1970, pp. 41–46) for some detailed discussion of this terminology.

So the study compared eight cases of vaginal cancer to 32 matched referents to see if treatment with diethylstilbestrol was more common among mothers of the cases, and indeed it was. Among the mothers of the eight cases, seven had received DES during pregnancy. Among mothers of the 32 referents, none had received DES. The association between vaginal cancer and DES appears to be almost as strong as a relationship can be, though of course only eight cases have been observed. If a conventional test designed for use in a randomized experiment is used to compare cases and referents in terms of the frequency of exposure to DES, the difference is highly significant. However, experience with the first example, vitamin C and cancer, suggests caution here.

What should be concluded from the strong association observed between DES and vaginal cancer in eight cases and 32 matched referents? Unlike the case of vitamin C and cancer, it would be neither practical nor ethical to follow up with a randomized experiment. Could such a hypothetical experiment produce very different findings? That possibility can never be entirely ruled out. Still, it is possible to ask: How severe would the unseen problems in this study have to be to produce such a strong relationship if DES did not cause vaginal cancer? How far would the observational study have to depart from an experiment to produce such a relationship if DES were harmless? How does the small size of the case group, eight cases, affect these questions? Chapter 4 provides answers. As it turns out, only severe unseen problems and hidden biases, only dramatic departures from an experiment, could produce such a strong association in the absence of an effect of DES, the small sample size notwithstanding. In other words, this study is highly insensitive to hidden bias; its conclusions could be altered by dramatic biases, but not by small ones. This is by no means true of all observational studies. Chapter 4 concerns general methods for quantifying the sensitivity of findings to hidden biases, and it discusses the uses and limitations of sensitivity analyses.

## *Academic Achievement in Public and Catholic High Schools: Specific Responses to Specific Criticisms*

A current controversy in the United States concerns the effectiveness of public or state-run schools, particularly as compared to existing privately operated schools. The 1985 paper by Hoffer, Greely, and Coleman is one of a series of observational studies of this question. They used data from the High School and Beyond Study (HSB), which includes a survey of US high-school students as sophomores with follow-up in their senior year. The HSB study provided standardized achievement test scores in several areas in sophomore and senior years, and included follow-up of students who dropped out of school, so as these things go, it is a rather complete and attractive source of data. Hoffer, Greely, and Coleman (1985) begin with a list of six objections made to their earlier studies, which had compared achievement test scores in public and Catholic schools, concluding that "... Catholic high schools are more effective than public high schools." As an illustration, objection #3 states: "Catholic schools seem to have an effect because they eliminate their disciplinary problems by expelling them from the school." The idea here is that Catholic schools eliminate difficult students while the public schools do not, so the students who remain in Catholic schools would be more likely to perform well even if there were no difference in the effectiveness of the two types of schools.

Criticism is enormously important to observational studies. The quality of the criticism offered in a particular field is intimately connected with the quality of the studies conducted in that field. Quality is not quantity, nor is harshness quality. What is scientifically plausible must be distinguished from what is just logically possible (Gastwirth, Krieger and Rosenbaum 1997). Cochran (1965, §5) argues that the first critic of an observational study should be its author:

> When summarizing the results of a study that shows an association consistent with the causal hypothesis, the investigator should always list and discuss all alternative explanations of his results (including different hypotheses and biases in the results) that occur to him. This advice may sound trite, but in practice is often neglected.

Criticisms of observational studies are of two kinds, the tangible and the dismissive, objection #3 being of the tangible kind. A tangible criticism is a specific and plausible alternative interpretation of the available data; indeed, a tangible criticism is itself a scientific theory, itself capable of empirical investigation. Bross (1960) writes:

> ... a critic who objects to a bias in the design or a failure to control some established factor is, in fact, raising a counterhypothesis ... [and] has the responsibility for showing [it] is tenable. In doing so, he operates under the same ground rules as the proponent ... : When a critic has shown that his counterhypothesis is tenable, his job is done, while at this point the proponent's job is just beginning. A proponent's job is not finished as long as there is a tenable hypothesis that rivals the one he asserts.

On the second page of his *The Design of Experiments*, Fisher (1935) described dismissive criticism as he argued that a theory of experimental design is needed:

> This type of criticism is usually made by what I might call a heavyweight *authority*. Prolonged experience, or at least the long possession of a scientific reputation, is almost a pre-requisite for developing successfully this line of attack. Technical details are seldom in evidence. The authoritative assertion: "His *controls* are *totally* inadequate" must have temporarily discredited many a promising line of work; and such an authoritarian method of judgement must surely continue, human nature being what it is, so long as theoretical notions of the principles of experimental design are lacking ... .

Dismissive criticism rests on the authority of the critic and is so broad and vague that its claims cannot be studied empirically. Judging the weight of evidence is inseparable from judging the criticisms that have been or can be raised.

Concerning objection #3, Hoffer, Greely, and Coleman (1985) respond: "... the evidence from the HSB data, although indirect, does not support this objection. Among students who reported that they had been suspended during their sophomore year, those in the Catholic sector were more likely to be in the same school as seniors than those in the public sector (63 percent to 56 percent)." In other words, difficult students, or at any rate students who were suspended, remained in Catholic school more often, not less often, than in public schools. This response to objection #3, though not decisive, does gives one pause.

Successful criticism of an observational study points to ambiguity in evidence or argument, and then points to methods for removing the ambiguity. Efforts to resolve an ambiguity are sometimes undermined by efforts to win an argument. Popper (1994, p. 44) writes:

> Serious critical discussions are always difficult ... Many participants in a rational, that is, a critical, discussion find it particularly difficult to unlearn what their instincts seem to teach them (and what they are taught, incidently, by every debating society): that is, to win. For what they have to learn is that victory in debate is nothing, while even the slightest clarification of one's problem—even the smallest contribution made towards a clearer understanding of one's own position or that of one's opponent—is a great success. A discussion which you win but which fails to help you change or to clarify your mind at least a little should be regarded as a sheer loss.

## 1.3 Purpose of This Book

Scientific evidence is commonly and properly greeted with objections, skepticism, and doubt. Some objections come from those who simply do not like the conclusions, but setting aside such unscientific reactions, responsible scientists are responsibly skeptical. We look for failures of observation, gaps in reasoning, alternative interpretations. We compare new evidence with past evidence. This skepticism is itself scrutinized. Skepticism must be justified, defended. One needs "grounds for doubt," in Wittgenstein's (1969, §122) phrase. The grounds for doubt are themselves challenged. Objections bring forth counterobjections and more evidence. As time passes, arguments on one side or the other become strained, fewer scientists are willing to offer them, and the arguments on that side come increasingly from individuals who seem to have some stake in the outcome. In this way, questions are settled.

Scientific questions are not settled on a particular date by a single event, nor are they settled irrevocably. We speak of the weight of evidence. Eventually, the weight is such that critics can no longer lift it, or are too weary to try. Overwhelming evidence is evidence that overwhelms responsible critics.

Experiments are better than observational studies because there are fewer grounds for doubt. The ideal experiment would leave few grounds for doubt, and at times this ideal is nearly achieved, particularly in the laboratory. Experiments often settle questions faster.

Despite this, experiments are not feasible in some settings. At times, observational studies have produced overwhelming evidence, as compelling as any in science, but at other times, observational data have misled investigators to advocate harmful policies or ineffective treatments.

A statistical theory of observational studies is a framework and a set of tools that provide measures of the weight of evidence. The purpose of this book is to give an account of statistical principles and methods for the design and analysis of observational studies. An adequate account must relate observational studies to controlled experiments, showing how uncertainty about treatment effects is greater in the absence of randomization. Analytical adjustments are common in observational studies, and the account should indicate what adjustments can and cannot do. A large literature offers many devices to detect hidden biases in observational studies, for instance, the use of several control groups, and the account must show how such devices work and when they may be expected to succeed or fail. Even when it is not possible to reduce or dispel uncertainty, it is possible to be careful in discussing its magnitude. That is, even when it is not possible to remove bias through adjustment or to detect bias through careful design, it is nonetheless possible to give quantitative expression to the magnitude of uncertainties about bias, a technique called *sensitivity analysis.* The account must indicate what can and cannot be done with a sensitivity analysis.

## 1.4 Bibliographic Notes

Most scientific fields that study human populations conduct observational studies. Many fields have developed a literature on the design, conduct, and interpretation of observational studies, often with little reference to related work in other fields. It is not possible to do justice to these several literatures in a short bibliographic note. There follows a short and incomplete list of fine books that contain substantial general discussions of the methodology used for observational studies in epidemiology, public program evaluation, or the social sciences. A shared goal in these diverse works is evaluation of treatments, exposures, programs, or policies from nonexperimental data. The list is followed by references cited in Chapter 1.

### *Some Books and a Few Papers*

Angrist, J. D. and Krueger, A. B. (1999) Empirical strategies in labor economics. In: *Handbook of Labor Economics*, O. Ashenfelter and D. Card, eds., Volume 3A, Chapter 23, New York: Elsevier.

Ashenfelter, O., ed. (2000) *Labor Economics*. New York: Worth.

Becker, H. S. (1997) *Tricks of the Trade*. Chicago: University of Chicago Press.

Blaug, M. (1980) *The Methodology of Economics*. New York: Cambridge University Press.

Breslow, N. and Day, N. (1980, 1987) *Statistical Methods in Cancer Research*, Volumes 1 and 2. Lyon, France: International Agency for Research on Cancer.

Campbell, D. T. (1988) *Methodology and Epistemology for Social Science: Selected Papers*. Chicago: University of Chicago Press, pp. 315—333.

Campbell, D. and Stanley, J. (1963) *Experimental and Quasi-Experimental Design for Research*. Chicago: Rand McNally.

Chamberlain, G. (1984) Panel data. In: *Handbook of Econometrics*, Chapter 22, Volume 2, Z. Griliches and M. D. Intriligator, eds., New York: Elsevier.

Cochran, W. G. (1965) The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society*, Series **A**, **128**, 134- -155.

Cochran, W. (1983) *Planning and Analysis of Observational Studies*. New York: Wiley.

Cook, T. D. and Campbell, D. C. (1979) *Quasi-Experimentation*. Chicago: Rand McNally.

Cook, T. D., Campbell, D. T., and Peracchio, L. (1990) Quasi-experimentation. In: *Handbook of Industrial and Organizational Psychology*, M. Dunnette and L. Hough, eds., Palo Alto, CA: Consulting Psychologists Press, Chapter 9, pp. 491—576.

Cook, T. D. and Shadish, W. R. (1994) Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology*, **45**, 545—580.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959) Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, **22**, 173—203.

Cox, D. R. (1992) Causality: Some statistical aspects. *Journal of the Royal Statistical Society*, Series **A**, **155**, 291--301.

Elwood, J. M. (1988) *Causal Relationships in Medicine.* New York: Oxford University Press.

Emerson, R. M. (1981) Observational field work. *Annual Review of Sociology*, **7**, 351—378.

Freedman, D. (1997) From association to causation via regression. *Advances in Applied Mathematics*, **18**, 59—110.

Friedman, M. (1953) *Essays in Positive Economics.* Chicago: University of Chicago Press.

Gastwirth, J. (1988) *Statistical Reasoning in Law and Public Policy.* New York: Academic Press.

Gordis, L. (2000) *Epidemiology* (Second Edition) Philadelphia: Saunders.

Greenhouse, S. (1982) Jerome Cornfield's contributions to epidemiology. *Biometrics*, **28**, Supplement, 33–46.

Heckman, J. J. (2001) Micro data, heterogeneity, and the evaluation of public policy: The Nobel lecture. *Journal of Political Economy*, **109**, 673–748.

Hill, A. B. (1965) The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, **58**, 295—300.

Holland, P. (1986) Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, **81**, 945–970.

Kelsey, J., Whittemore, A., Evans, A., and Thompson. W. (1996). *Methods in Observational Epidemiology.* New York: Oxford University Press.

Khoury, M. J., Cohen, B. H., and Beaty, T. H. (1993) *Fundamentals of Genetic Epidemiology.* New York: Oxford University Press.

Kish, L. (1987) *Statistical Design for Research.* New York: Wiley.

Lilienfeld, A. and Lilienfeld, D. E. (1980) *Foundations of Epidemiology.* New York: Oxford University Press.

Lilienfeld, D. E. and Stolley, P. D. (1994) *Foundations of Epidemiology.* New York: Oxford University Press.

Lipsey, M. W. and Cordray, D. S. (2000) Evaluation methods for social intervention. *Annual Review of Psychology*, **51**, 345—375.

Little, R. J. and Rubin, D. B. (2000) Causal effects in clinical and epidemiological studies via potential outcomes. *Annual Review of Public Health*, **21**, 121– 145.

Maclure, M. and Mittleman, M. A. (2000) Should we use a case-crossover design? *Annual Review of Public Health*, **21**, 193 –221.

MacMahon, B. and Pugh, T. (1970) *Epidemiology.* Boston: Little, Brown.

MacMahon, B. and Trichopoulos, D. (1996) *Epidemiology.* Boston: Little, Brown.

Manski, C. (1995) *Identification Problems in the Social Sciences.* Cambridge, MA: Harvard University Press.

Mantel, N. and Haenszel, W. (1959) Statistical aspects of retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719–748.

Meyer, B. D. (1995) Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics*, **13**, 151—161.

Meyer, M. and Fienberg, S., eds. (1992) *Assessing Evaluation Studies: The Case of Bilingual Education Strategies.* Washington, DC: National Academy Press.

Miettinen, O. (1985) *Theoretical Epidemiology.* New York: Wiley.

Pearl, J. (2000) *Causality: Models, Reasoning, Inference.* New York: Cambridge University Press.

Reichardt, C. S. (2000) A typology of strategies for ruling out threats to validity. In: *Research Design: Donald Campbell's Legacy*, L. Brickman, ed., Thousand Oaks, CA: Sage, Volume 2, pp., 89–115.

Reiter, J. (2000) Using statistics to determine causal relationships. *American Mathematical Monthly*, **107**, 24—32.

Robins, J. M. (1999) Association, causation, and marginal structural models. *Synthese*, **121**, 151—179.

Robins, J., Blevins, D., Ritter, G., and Wulfsohn, M. (1992) G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology,* **3**, 319—336.

Rosenthal, R. and Rosnow, R., eds. (1969) *Artifact in Behavioral Research.* New York: Academic.

Rosenzweig, M. R. and Wolpin, K. I. (2000) Natural "natural experiments" in economics. *Journal of Economic Literature*, **38**, 827—874.

Rosnow, R. L. and Rosenthal, R. (1997) *People Studying People: Artifacts and Ethics in Behavioral Research.* New York: W. H. Freeman.

Rossi, P., Freeman, H., and Lipsey, M. W. (1999) *Evaluation.* Beverly Hills, CA: Sage.

Rothman, K. and Greenland, S. (1998) *Modern Epidemiology.* Philadelphia: Lippincott-Raven.

Rubin, D. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688- 701.

Schlesselman, J. (1982) *Case-Control Studies.* New York: Oxford University Press.

Schulte, P. A. and Perera, F. (1993) *Molecular Epidemiology: Principles and Practices.* New York: Academic.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton-Mifflin.

Shafer, G. (1996) *The Art of Causal Conjecture.* Cambridge, MA: MIT Press.

Sobel, M. (1995) Causal inference in the social and behavioral sciences. In: *Handbook of Statistical Modelling for the Social and Behavioral Sciences*, G. Arminger, C. Clogg, and M. Sobel, eds., New York: Plenum, 1—38.

Steenland, K., ed. (1993) *Case Studies in Occupational Epidemiology.* New York: Oxford University Press.

Strom, B. (2000) *Pharmacoepidemiology.* New York: Wiley.

Suchman, E. (1967) *Evaluation Research.* New York: Sage.

Susser, M. (1973) *Causal Thinking in the Health Sciences: Concepts and Strategies in Epidemiology.* New York: Oxford University Press.

Susser, M. (1987) *Epidemiology, Health and Society: Selected Papers.* New York: Oxford University Press.

Tufte, E., ed. (1970) *The Quantitative Analysis of Social Problems.* Reading, MA: Addison-Wesley.

Weiss, C. (1997) *Evaluation.* Englewood Cliffs, NJ: Prentice-Hall.

Weiss, N. S. (1996) *Clinical Epidemiology.* New York: Oxford University Press.

Willett, W. (1998) *Nutritional Epidemiology.* New York: Oxford University Press.

Winship, C. and Morgan, S. L. (1999) The estimation of causal effects from observational data. *Annual Review of Sociology*, **25**, 659—706.

Zellner, A. (1968) *Readings in Economic Statistics and Econometrics.* Boston: Little, Brown.

## 1.5 References

Bross, I. D. J. (1960) Statistical criticism. *Cancer*, **13**, 394–400. Reprinted in: *The Quantitative Analysis of Social Problems*, E. Tufte, ed., Reading, MA: Addison-Wesley, pp. 97–108.

Cameron, E. and Pauling, L. (1976) Supplemental ascorbate in the supportive treatment of cancer: Prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Sciences* (USA), **73**, 3685–3689.

Chalmers, T., Block, J., and Lee, S. (1970) Controlled studies in clinical cancer research. *New England Journal of Medicine*, **287**, 75–78.

Cochran, W.G. (1965) The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society*, Series A, **128**, 134–155. Reprinted in *Readings in Economic Statistics and Econometrics*, A. Zellner, ed., 1968, Boston: Little Brown, pp. 11–36.

Dehejia, R. H. and Wahba, S. (1999) Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association,* **94**, 1053—1062.

Doll, R. and Hill, A. (1966) Mortality of British doctors in relation to smoking: Observations on coronary thrombosis. In: *Epidemiological Approaches to the Study of Cancer and Other Chronic Diseases*, W. Haenszel, ed., U.S. National Cancer Institute Monograph 19, Washington, DC: US Department of Health, Education, and Welfare, pp. 205–268.

Fisher, R.A. (1935, 1949) *The Design of Experiments.* Edinburgh: Oliver & Boyd.

Fraker, T. and Maynard, R. (1987) The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*, **22**, 194–227.

Friedlander, D. and Robins, P. K. (1995) Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *American Economic Review*, **85**, 923—937.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1997) Hypotheticals and hypotheses. *American Statistician*, **51**, 120—121.

Herbst, A., Ulfelder, H., and Poskanzer, D. (1971) Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women. *New England Journal of Medicine*, **284**, 878–881.

Hoffer, T., Greeley, A., and Coleman, J. (1985) Achievement growth in public and Catholic schools. *Sociology of Education*, **58**, 74–97.

LaLonde, R. (1986) Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, **76**, 604–620.

Meier, P. (1972) The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In: *Statistics: A Guide to the Unknown*, J. Tanur, ed., San Francisco: Holden-Day, pp. 2–13.

Moertel, C., Fleming, T., Creagan, E., Rubin, J., O'Connell, M., and Ames, M. (1985) High-dose vitamin C versus placebo in the treatment of patients with advanced cancer who have had no prior chemotherapy: A randomized double-blind comparison. *New England Journal of Medicine*, **312**, 137–141.

Popper, K. (1959) *The Logic of Scientific Discovery*. New York: Harper & Row.

Popper, K. (1994) *The Myth of the Framework*. New York: Routledge.

United States Surgeon General's Advisory Committee Report (1964) *Smoking and Health*. Washington, DC: US Department of Health, Education and Welfare.

Wittgenstein, L. (1969) *On Certainty*. New York: Harper & Row.

Zwick, R. (1991) Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, **3**, 10–16.

# 2
# Randomized Experiments

## 2.1 Introduction and Example: A Randomized Clinical Trial

Observational studies and controlled experiments have the same goal, inference about treatment effects, but random assignment of treatments is present only in experiments. This chapter reviews the role of randomization in experiments, and so prepares for discussion of observational studies in later chapters. A theory of observational studies must have a clear view of the role of randomization, so it can have an equally clear view of the consequences of its absence. Sections 2.1 and 2.2 give two examples: a large controlled clinical trial, and then a small but famous example due to Sir Ronald Fisher, who is usually credited with the invention of randomization, which he called the "reasoned basis for inference" in experiments. Later sections discuss the meaning of this phrase, that is, the link between randomization and statistical methods. Most of the material in this chapter is quite old.

### *Randomized Trial of Coronary Surgery*

The US Veterans Administration (Murphy et al. 1977) conducted a randomized controlled experiment comparing coronary artery bypass surgery with medical therapy as treatments for coronary artery disease. Bypass surgery is an attempt to repair the arteries that supply blood to the heart, arteries that have been narrowed by fatty deposits. In bypass surgery, a

TABLE 2.1. Base-Line Comparison of Coronary Patients in the Veterans Administration Randomized Trial.

| Covariate | Medical<br>% | Surgical<br>% |
|---|---|---|
| New York Heart Association Class II & III | 94.2 | 95.4 |
| History of myocardial infarction (MI) | 59.3 | 64.0 |
| Definite or possible MI based on electrocardiogram | 36.1 | 40.5 |
| Duration of chest pain > 25 months | 50.0 | 51.8 |
| History of hypertension | 30.0 | 27.6 |
| History of congestive heart failure | 8.4 | 5.2 |
| History of cerebral vascular episode | 3.2 | 2.1 |
| History of diabetes | 12.9 | 12.2 |
| Cardiothoracic ratio > 0.49 | 10.4 | 12.2 |
| Serum cholesterol > 249 mg/100 ml | 31.6 | 20.6 |

bypass or bridge is formed around a blockage in a coronary artery. In contrast, medical therapy uses drugs to enhance the flow of blood through narrowed arteries. The study involved randomly assigning 596 patients at 13 Veterans Administration hospitals, of whom 286 received surgery and 310 received drug treatments. The random assignment of a treatment for each patient was determined by a central office after the patient had been admitted into the trial.

Table 2.1 is taken from their study. It compares the medical and surgical treatment groups in terms of 10 important characteristics of patients measured at "base-line," that is, prior to the start of treatment. A variable measured prior to the start of treatment is called a *covariate*. Similar tables appear in reports of most clinical trials.

Table 2.1 shows the two groups of patients were similar in many important ways prior to the start of treatment, so that comparable groups were being compared. When the percentages for medical and surgical are compared, the difference is not significant at the 0.05 level for nine of the variables in Table 2.1, but is significant for serum cholesterol. This is in line with what one would expect from 10 significance tests if the only differences were due to chance, that is, due to the choice of random numbers used in assigning treatments.

For us, Table 2.1 is important for two reasons. First, it is an example showing that randomization tends to produce relatively comparable or balanced treatment groups in large experiments. The second point is separate and more important. The 10 covariates in Table 2.1 were not used in assigning treatments. There was no deliberate balancing of these variables. Rather the balance we see was produced by the random assignment, which made no use of the variables themselves. This gives us some reason to hope and expect that other variables, not measured, are similarly balanced. Indeed, as shown shortly, statistical theory supports this expectation. Had the trial not used random assignment, had it instead assigned patients one at a time to balance these 10 covariates, then the balance might well have been better than in Table 2.1, but there would have been no basis for expecting other unmeasured variables to be similarly balanced.

The VA study compared survival in the two groups three years after treatment. Survival in the medical group was 87% and in the surgical group 88%, both with a standard error of 2%. The 1% difference in mortality was not significant. Evidently, when comparable groups of patients received medical and surgical treatment at the VA hospitals, the outcomes were quite similar.

The statement that randomization tends to balance covariates is at best imprecise; taken too literally, it is misleading. For instance, in Table 2.1, the groups do differ slightly in terms of serum cholesterol. Presumably there are other variables, not measured, exhibiting imbalances similar to if not greater than that for serum cholesterol. What is precisely true is that random assignment of treatments can produce some imbalances by chance, but common statistical methods, properly used, suffice to address the uncertainty introduced by these chance imbalances. To this subject, we now turn.

## 2.2 The Lady Tasting Tea

"A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup," or so begins the second chapter of Sir Ronald Fisher's (1935, 1949) book *The Design of Experiments*, which introduced the formal properties of randomization. This example is part of the tradition of statistics, and in addition it was well selected by Fisher to illustrate key points. He continues:

> Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgement in a random order. The subject has been told in advance of what the test will consist, namely that she

> will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such a manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received.

Fisher then asks what would be expected if the Lady was "without any faculty of discrimination," that is, if she makes no changes at all in her judgments in response to changes in the order in which tea and milk are added to the cups. To change her judgments is to have some faculty of discrimination, however slight. So suppose for the moment that she cannot discriminate at all, that she gives the same judgments no matter which four cups receive milk first. Then it is only by accident or chance that she correctly identifies the four cups in which milk was added first. Since there are $\binom{8}{4} = 70$ possible divisions of the eight cups into two groups of four, and randomization has ensured that these are equally probable, the chance of this accident is 1/70. In other words, the probability that the random ordering of the cups will yield perfect agreement with the Lady's fixed judgments is 1/70. If the Lady correctly classified the cups, this probability, $0.014 = 1/70$, is the significance level for testing the null hypothesis that she is without the ability to discriminate.

Fisher goes on to describe randomization as the "reasoned basis" for inference and "the physical basis of the validity of the test"; indeed, these phrases appear in section headings and are clearly important to Fisher. He explains:

> We have now to examine the physical conditions of the experimental technique needed to justify the assumption that, if discrimination of the kind under test is absent, the result of the experiment will be wholly governed by the laws of chance ... .
>
> It is [not sufficient] to insist that "all the cups must be exactly alike" in every respect except that to be tested. For this is a totally impossible requirement in our example, and equally in all other forms of experimentation ... .
>
> The element in the experimental procedure which contains the essential safeguard is that the two modifications of the test beverage are to be prepared "in random order." This, in fact, is the only point in the experimental procedure in which the laws of chance, which are to be in exclusive control of our frequency distribution, have been explicitly introduced.

Fisher discusses this example for 15 pages, though its formal aspects are elementary and occupy only a part of a paragraph. He is determined to establish that randomization has justified or grounded a particular inference, formed its "reasoned basis," a basis that would be lacking had the same pattern of responses, the same data, been observed in the absence of randomization.

The example serves Fisher's purpose well. The Lady is not a sample from a population of Ladies, and even if one could imagine that she was, there is but one Lady in the experiment and the hypothesis concerns her alone. Her eight judgments are not independent observations, not least because the rules require a split into four and four. Later cups differ from earlier ones, for by cup number five, the Lady has surely tasted one with milk first and one with tea first. There is no way to construe, or perhaps misconstrue, the data from this experiment as a sample from a population, or as a series of independent and identical replicates. And yet, Fisher's inference is justified, because the only probability distribution used in the inference is the one created by the experimenter.

What are the key elements in Fisher's argument? First, experiments do not require, indeed cannot reasonably require, that experimental units be homogeneous, without variability in their responses. Homogeneous experimental units are not a realistic description of factory operations, hospital patients, agricultural fields. Second, experiments do not require, indeed, cannot reasonably require, that experimental units be a random sample from a population of units. Random samples of experimental units are not the reality of the industrial laboratory, the clinical trial, or the agricultural experiment. Third, for valid inferences about the effects of a treatment on the units included in an experiment, it is sufficient to require that treatments be allocated at random to experimental units—these units may be both heterogeneous in their responses and not a sample from a population. Fourth, probability enters the experiment only through the random assignment of treatments, a process controlled by the experimenter. A quantity that is not affected by the random assignment of treatments is a fixed quantity describing the units in the experiment.

The next section repeats Fisher's argument in more general terms.

## 2.3 Randomized Experiments

### *2.3.1 Units and Treatment Assignments*

There are $N$ units available for experimentation. A unit is an opportunity to apply or withhold the treatment. Often, a unit is a person who will receive either the treatment or the control as determined by the experimenter. However, it may happen that it is not possible to assign a treatment to a single person, so a group of people form a single unit, perhaps all children in a particular classroom or school. On the other hand, a single person may present several opportunities to apply different treatments, in which case each opportunity is a unit; see Problem 2. For instance, in §2.2, the one Lady yielded eight units.

The $N$ units are divided into $S$ strata or subclasses on the basis of covariates, that is, on the basis of characteristics measured prior to the assignment of treatments. The stratum to which a unit belongs is not affected by the treatment, since the strata are formed prior to treatment. There are $n_s$ units in stratum $s$ for $s = 1, \dots, S$, so $N = \sum n_s$.

Write $Z_{si} = 1$ if the $i$th unit in stratum $s$ receives the treatment and write $Z_{si} = 0$ if this unit receives the control. Write $m_s$ for the number of treated units in stratum $s$, so $m_s = \sum_{i=1}^{n_s} Z_{si}$, and $0 \le m_s \le n_s$. Finally, write $\mathbf{Z}$ for the $N$-dimensional column vector containing the $Z_{si}$ for all units in the lexical order; that is,

$$\mathbf{Z} = \begin{bmatrix} Z_{11} \\ Z_{12} \\ \vdots \\ Z_{1,n_1} \\ Z_{21} \\ \vdots \\ Z_{S,n_S} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_S \end{bmatrix}, \quad \text{where} \quad \mathbf{Z}_s = \begin{bmatrix} Z_{s1} \\ \vdots \\ Z_{s,n_s} \end{bmatrix}. \tag{2.1}$$

This notation covers several common situations. If no covariates are used to divide the units, then there is a single stratum containing all units, so $S = 1$. If $n_s = 2$ and $m_s = 1$ for $s = 1, \dots, S$, then there are $S$ pairs of units matched on the basis of covariates, each pair containing one treated unit and one control. The situation in which $n_s \ge 2$ and $m_s = 1$ for $s = 1, \dots, S$ is called matching with multiple controls. In this case there are $S$ matched sets, each containing one treated unit and one or more controls.

The case of a single stratum, that is $S = 1$, is sufficiently common and important to justify slight modifications in notation. When there is only a single stratum the subscript $s$ is dropped, so $Z_i$ is written in place of $Z_{1i}$. The same convention applies to other quantities that have subscripts $s$ and $i$.

### 2.3.2 Several Methods of Assigning Treatments at Random

In a *randomized experiment*, the experimenter determines the assignment of treatments to units, that is the value of $\mathbf{Z}$, using a known random mechanism such as a table of random numbers. To say that the mechanism is known is to say that the distribution of the random variable $\mathbf{Z}$ is known because it was created by the experimenter. One requirement is placed on this random mechanism, namely, that, before treatments are assigned, every unit has a nonzero chance of receiving both the treatment and the control, or formally that $0 < \text{prob}(Z_{si} = 1) < 1$ for $s = 1, \ldots, S$ and $i = 1, \ldots, n_s$. Write $\Omega_0$ for the set containing all possible values of $\mathbf{Z}$, that is, all values of $\mathbf{Z}$ which are given nonzero probability by the mechanism.

In practice, many different random mechanisms have been used to determine $\mathbf{Z}$. The simplest assigns treatments independently to different units, taking $\text{prob}(Z_{si} = 1) = 1/2$ for all $s$, $i$. This method was used in the Veterans Administration experiment on coronary artery surgery in §2.1. In this case, $\Omega_0$ is the set containing $2^N$ possible values of $\mathbf{Z}$, namely, all $N$-tuples of zeros and ones, and every assignment in $\Omega_0$ has the same probability; that is, $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/2^N$ for all $\mathbf{z} \in \Omega_0$. The number of elements in a set $A$ is written $|A|$, so in this case $|\Omega_0| = 2^N$. This mechanism has the peculiar property that there is a nonzero probability that all units will be assigned to the same treatment, though this probability is extremely small when $N$ is moderately large. From a practical point of view, a more important problem with this mechanism arises when $S$ is fairly large compared to $N$. In this case, the mechanism may give a high probability to the set of treatment assignments in which all units in some stratum receive the same treatment. If the strata were types of patients in a clinical trial, this would mean that all patients of some type received the same treatment. If the strata were schools in an educational experiment, it would mean that all children in some school received the same treatment. Other assignment mechanisms avoid this possibility.

The most commonly used assignment mechanism fixes the number $m_s$ of treated subjects in stratum $s$. In other words, the only assignments $\mathbf{Z}$ with nonzero probability are those with $m_s$ treated subjects in stratum $s$ for $s = 1, \ldots, S$. If $m_s$ is chosen sensibly, this avoids the problem mentioned in the previous paragraph. For instance, if $n_s$ is required to be even and $m_s$ is required to equal $n_s/2$ for each $s$, then half the units in each stratum receive the treatment and half receive the control, so the final treated and control groups are exactly balanced in the sense that they contain the same number of units from each stratum.

When $m_s$ is fixed in this way, let $\Omega$ be the set containing the $K = \prod_{s=1}^{S} \binom{n_s}{m_s}$ possible treatment assignments $\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_S \end{bmatrix}$ in which $\mathbf{z}_s$ is an $n_s$-tuple with $m_s$ ones and $n_s - m_s$ zeros for $s = 1, \ldots, S$. In the most common assignment mechanism, each of these $K$ possible assignments is given the same probability, $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K$ all $\mathbf{z} \in \Omega$. This type of randomized experiment, with equal probabilities and fixed $m_s$, will be called a *uniform randomized experiment*. When there is but a single stratum, $S = 1$, it has traditionally been called a *completely randomized experiment*, but when there are two or more strata, $S \geq 2$, it has been called a *randomized block experiment*. If the strata each contain two units, $n_s = 2$, and one receives the treatment, $m_s = 1$, then it has been called a *paired randomized experiment*.

As a small illustration, consider a uniform randomized experiment with two strata, $S = 2$, four units in the first stratum, $n_1 = 4$, and two in the second, $n_2 = 2$, and $N = n_1 + n_2 = 6$ units in total. Half of the units in each stratum receive the treatment, so $m_1 = 2$ and $m_2 = 1$. There are $K = 12$ possible treatment assignments $\mathbf{z} = (z_{11}, z_{12}, z_{13}, z_{14}, z_{21}, z_{22})^T$ contained in the set $\Omega$, and each has probability 1/12. So $\Omega$ is the following set of $K = 12$ vectors $\mathbf{z}$ of dimension $N = 6$ with binary coordinates such that $2 = z_{11} + z_{12} + z_{13} + z_{14}$ and $1 = z_{21} + z_{22}$.

$$\Omega = \left\{ \begin{array}{cccccc} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \\ \\ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \end{array} \right\}.$$

The following proposition is often useful. It says that in a uniform randomized experiment, the assignments in different strata are independent of each other. For the elementary proof, see Problem 3.

**Proposition 1** *In a uniform randomized experiment, the* $\mathbf{Z}_1, \ldots, \mathbf{Z}_S$ *are mutually independent, and* $\text{prob}(\mathbf{Z}_s = \mathbf{z}_s) = 1/\binom{n_s}{m_s}$ *for each* $n_s$*-tuple* $\mathbf{z}_s$ *containing* $m_s$ *ones and* $n_s - m_s$ *zeros.*

The uniform randomized designs are by far the most common randomized experiments involving two treatments, but others are also used, particularly in clinical trials. It is useful to mention one of these methods of randomization to underscore the point that randomized experiments need not give every treatment assignment $\mathbf{z} \in \Omega_0$ the same probability. A distinguishing feature of many clinical trials is that the units are patients who arrive for treatment over a period of months or years. As a result, the number $n_s$ of people who will fall in stratum $s$ will not be known at the start of the experiment, so a randomized block experiment is not possible. Efron (1971) proposed the following method. Fix a probability $p$ with $1/2 < p < 1$. When the $i$th patient belonging to stratum $s$ first arrives, calculate a current measure of imbalance in stratum $s$, $\text{IMBAL}_{si}$, defined to be the number of patients so far assigned to treatment in this stratum minus the number so far assigned to control. It is easy to check that

$\text{IMBAL}_{si} = \sum_{j=1}^{i-1}(2Z_{sj} - 1)$. If $\text{IMBAL}_{si} = 0$, assign the new patient to treatment or control each with probability 1/2. If $\text{IMBAL}_{si} < 0$, so there are too few treated patients in this stratum, then assign the new patient to treatment with probability $p$ and to control with probability $1 - p$. If $\text{IMBAL}_{si} > 0$, so there are too many treated patients, then assign the new patient to treatment with probability $1 - p$ and to control with probability $p$. Efron examines various aspects of this method. In particular, he shows that it is much better than independent assignment in producing balanced treated and control groups, that is, treated and control groups with similar numbers of patients from each stratum. He also examines potential biases due to the experimenter's knowledge of $\text{IMBAL}_{si}$. Zelen (1974) surveys a number of related methods with similar objectives.

## 2.4 Testing the Hypothesis of No Treatment Effect

### *2.4.1 The Distribution of a Test Statistic When the Treatment Is Without Effect*

In the theory of experimental design, a special place is given to the test of the hypothesis that the treatment is entirely without effect. The reason is that, in a randomized experiment, this test may be performed virtually without assumptions of any kind, that is, relying just on the random assignment of treatments. Fisher discussed the Lady and her tea with such care to demonstrate this. Other activities, such as estimating treatment effects or building confidence intervals, do require some assumptions, often innocuous assumptions, but assumptions nonetheless. The contribution of randomization to formal inference is most clear when expressed in terms of the test of no effect. Does this mean that such tests are of greater practical importance than point or interval estimates? Certainly not. It is simply that the theory of such tests is less cluttered, and so it sets randomized and nonrandomized studies in sharper contrast. The important point is that, in the absence of difficulties such as noncompliance or loss to follow-up, assumptions play a minor role in randomized experiments, and no role at all in randomization tests of the hypothesis of no effect. In contrast, inference in a nonrandomized experiment requires assumptions that are not at all innocuous. So let us follow Fisher and develop this point with care.

Each unit exhibits a response that is observed some time after treatment. To say that the treatment has no effect on this response is to say that each unit would exhibit the same value of the response whether assigned to treatment or control. If the treatment has no effect on a patient's survival, then the patient would live the same number of months under treatment or under control. This is the definition of "no effect." If changing the treatment assigned to a unit changed that unit's response, then certainly the treatment has at least some effect. If a patient would live one more month under treatment than under control, then the treatment has some effect on that patient.

In the traditional development of randomization inference, chance and probability enter only through the random assignment of treatments, that is, through the known mechanism that selects the treatment assignment $\mathbf{Z}$ from $\Omega$. The only random quantities are $\mathbf{Z}$ and quantities that depend on $\mathbf{Z}$. When the treatment is without effect, the response of a unit is fixed, in the sense that this response would not change if a different treatment assignment $\mathbf{Z}$ were selected from $\Omega$. Again, this is simply what it means for a treatment to be without effect. When testing the null hypothesis of no effect, the response of the $i$th unit in stratum $s$ is written $r_{si}$ and the $N$-tuple of responses for all $N$ units is written $\mathbf{r}$. The lowercase notation for $r_{si}$ emphasizes that, under the null hypothesis, $r_{si}$ is a fixed quantity and not a random variable. Later on, when discussing treatments with effects, a different notation is needed.

A test statistic $t(\mathbf{Z}, \mathbf{r})$ is a quantity computed from the treatment assignment $\mathbf{Z}$ and the response $\mathbf{r}$. For instance, the treated-minus-control difference in sample means is the test statistic

$$t(\mathbf{Z}, \mathbf{r}) = \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{r}}{\mathbf{Z}^{\mathrm{T}}\mathbf{1}} - \frac{(\mathbf{1}-\mathbf{Z})^{\mathrm{T}}\mathbf{r}}{(\mathbf{1}-\mathbf{Z})^{\mathrm{T}}\mathbf{1}}, \tag{2.2}$$

where $\mathbf{1}$ is an $N$-tuple of 1s. Other statistics are discussed shortly.

Given any test statistic $t(\mathbf{Z}, \mathbf{r})$, the task is to compute a significance level for a test that rejects the null hypothesis of no treatment effect when $t(\mathbf{Z}, \mathbf{r})$ is large. More precisely:

(i) The null hypotheses of no effect is tentatively assumed to hold, so $\mathbf{r}$ is fixed.

(ii) A treatment assignment $\mathbf{Z}$ has been selected from $\Omega$ using a known random mechanism.

(iii) The observed value, say $T$, of the test statistic $t(\mathbf{Z}, \mathbf{r})$ has been calculated.

(iv) We seek the probability of a value of the test statistic as large or larger than that observed if the null hypothesis were true.

The significance level is simply the sum of the randomization probabilities of assignments $\mathbf{z} \in \Omega$ that lead to values of $t(\mathbf{z}, \mathbf{r})$ greater than or equal to the observed value $T$, namely,

$$\operatorname{prob}\{t(\mathbf{Z}, \mathbf{r}) \geq T\} = \sum_{\mathbf{z} \in \Omega} [t(\mathbf{z}, \mathbf{r}) \geq T] \cdot \operatorname{prob}(\mathbf{Z} = \mathbf{z}), \tag{2.3}$$

$$\text{where } [\text{event}] = \begin{cases} 1 & \text{if event occurs,} \\ 0 & \text{otherwise,} \end{cases} \tag{2.4}$$

and $\text{prob}(\mathbf{Z} = \mathbf{z})$ is determined by the known random mechanism that assigned treatments. This is a direct calculation, though not always a straightforward one when $\Omega$ is extremely large.

In the case of a uniform randomized experiment, there is a simpler expression for the significance level (2.3) since $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K = 1/|\Omega|$. It is the proportion of treatment assignments $\mathbf{z} \in \Omega$ giving values of the test statistic $t(\mathbf{z}, \mathbf{r})$ greater than or equal to $T$, namely,

$$\text{prob}\{t(\mathbf{Z}, \mathbf{r}) \geq T\} = \frac{|\{\mathbf{z} \in \Omega : t(\mathbf{z}, \mathbf{r}) \geq T\}|}{K}. \tag{2.5}$$

### *2.4.2 More Tea*

To illustrate, consider again Fisher's example of the Lady who tastes $N = 8$ cups of tea, all in a single stratum, so $S = 1$. A treatment assignment is an 8-tuple containing four 1s and four 0s. For instance, the assignment $\mathbf{Z} = (1, 0, 0, 1, 1, 0, 0, 1)^{\mathrm{T}}$ would signify that cups 1, 4, 5, and 8 had milk added first and the other cups had tea added first. The set of treatment assignments $\Omega$ contains all possible 8-tuples containing four 1s and four 0s, so $\Omega$ contains $|\Omega| = K = \binom{8}{4} = 70$ such 8-tuples. The actual assignment was selected at random in the sense that $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K = 1/70$ for all $\mathbf{z} \in \Omega$. Notice that $\mathbf{z}^{\mathrm{T}}\mathbf{1} = 4$ for all $\mathbf{z} \in \Omega$.

The Lady's response for cup $i$ is either $r_i = 1$ signifying that she classifies this cup as milk first or $r_i = 0$ signifying that she classifies it as tea first. Then $\mathbf{r} = (r_1, \ldots, r_8)^{\mathrm{T}}$. Recall that she must classify exactly four cups as milk first, so $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 4$. The test statistic is the number of cups correctly identified, and this is written formally as $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}r + (\mathbf{1} - \mathbf{Z})^{\mathrm{T}}(\mathbf{1} - \mathbf{r}) = 2\mathbf{Z}^{\mathrm{T}}\mathbf{r}$, where the second equality follows from $\mathbf{1}^{\mathrm{T}}\mathbf{1} = 8$, $\mathbf{Z}^{\mathrm{T}}\mathbf{1} = 4$, and $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 4$. To make this illustration concrete, suppose that $\mathbf{r} = (1, 1, 0, 0, 0, 1, 1, 0)$, so the Lady classifies the first, second, sixth, and seventh cups as milk first. To say that the treatment has no effect is to say that she would give this classification no matter how milk was added to the cups, that is, no matter how treatments were assigned to cups. If changing the cups to which milk is added first changes her responses, then she is discerning something, and the treatment has some effect, however slight or erratic.

There is only one treatment assignment $\mathbf{z} \in \Omega$ leading to perfect agreement with the Lady's responses, namely, $\mathbf{z} = (1, 1, 0, 0, 0, 1, 1, 0)$, so if $t(\mathbf{Z}, \mathbf{r}) = 8$ the significance level (2.5) is $\text{prob}\{t(\mathbf{Z}, \mathbf{r}) \geq 8\} = 1/70$. This says that the chance of perfect agreement by accident is $1/70 = 0.014$, a small chance. In other words, if the treatment is without effect, the chance that a random assignment of treatments will just happen to produce perfect agreement is $1/70$.

It is not possible to have seven agreements since to err once is to err twice. How many assignments $\mathbf{z} \in \Omega$ lead to exactly $t(\mathbf{Z}, \mathbf{r}) =$ six agreements? One such assignment with six agreements is $\mathbf{z} = (1, 0, 1, 0, 0, 1, 1, 0)$. Starting with perfect agreement, $\mathbf{z} = (1,1,0,0,0,1,1,0)$, any one of the four 1s may be made a 0 and any of the four 0s may be made a 1, so there are $16 = 4 \times 4$ assignments with exactly $t(\mathbf{Z},\mathbf{r}) = 6$ agreements. Hence, there are 17 assignments leading to six or more agreements. With six agreements the significance level (2.5) is $\text{prob}\{t(\mathbf{Z},\mathbf{r}) \geq 6\} = 17/70 = 0.24$, no longer a small probability. It would not be surprising to see six or more agreements if the treatment were without effect—it happens by chance as frequently as seeing two heads when flipping two coins.

The key point deserves repeating. Probability enters the calculation only through the random assignment of treatments. The needed probability distribution is known, not assumed. The resulting significance level does not depend upon assumptions of any kind. If the same calculation were performed in a nonrandomized study, it would require an assumption that the distribution of treatment assignments, $\text{prob}(\mathbf{Z} = \mathbf{z})$, is some particular distribution, perhaps the assumption that all assignments are equally probable, $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K$. In a nonrandomized study, there may be little basis on which to ground or defend this assumption, it may be wrong, and it will certainly be open to responsible challenge and debate. In other words, the importance of the argument just considered is that it is one way of formally expressing the claim that randomized experiments are not open to certain challenges that can legitimately be made to nonrandomized studies.

### *2.4.3 Some Common Randomization Tests*

Many commonly used tests are randomization tests in that their significance levels can be calculated using (2.5), though the tests are sometimes derived in other ways as well. This section briefly recalls and reviews some of these tests. The purpose of the section is to provide a reference for these methods in a common terminology so they may be discussed and used at later stages. Though invented at different times, it is natural to see the methods as members of a few classes whose properties are similar, and this is done beginning in §2.4.4. In most cases, the methods described have various optimality properties which are not discussed here; see Cox (1970) for the optimality properties of the procedures for binary outcomes and Lehmann (1975) for optimality properties of the nonparametric procedures. In all cases, the experiment is the uniform randomized experiment in §2.3.2 with $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K$ for all $\mathbf{z} \in \Omega$.

*Fisher's exact test* for a $2 \times 2$ contingency table is, in fact, the test just used for the example of the Lady and her tea. Here, there is one stratum, $S = 1$; the outcome $r_i$ is *binary*, that is, $r_i = 1$ or $r_i = 0$; and the test statistic is the number of responses equal to 1 in the treated group, that is, $t(\mathbf{Z},\mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{r}$. The $2 \times 2$ contingency table records the values of $Z_i$ and $r_i$, as shown in Table 2.2 for Fisher's example. Notice that the marginal totals in this table are fixed by the structure of the experiment, because

TABLE 2.2. The 2 × 2 Table for Fisher's Exact Test for the Lady Tasting Tea.

| | | Response, $r_i$ | | |
|---|---|---|---|---|
| | | 1 | 0 | Total |
| Treatment or control, $Z_i$ | 1 | $\mathbf{Z}^T\mathbf{r}$ | $4 - \mathbf{Z}^T\mathbf{r}$ | 4 |
| | 0 | $4 - \mathbf{Z}^T\mathbf{r}$ | $\mathbf{Z}^T\mathbf{r}$ | 4 |
| | Total | 4 | 4 | 8 |

$N = 8$ cups, $\mathbf{1}^T\mathbf{r} = 4$ and $\mathbf{1}^T\mathbf{Z} = 4$ are fixed in this experiment. Under the hypothesis of no effect, the randomization distribution of the test statistic $\mathbf{Z}^T\mathbf{r}$ is the hypergeometric distribution. The usual *chi-square* test for a $2 \times 2$ table is an approximation to the randomization significance level when $N$ is large.

The *Mantel–Haenszel* (1959) *statistic* is the analogue of Fisher's exact test when there are two or more strata, $S \geq 2$, and the outcome $r_{si}$ is binary. It is extensively used in epidemiology and certain other fields. The data may be recorded in a $2 \times 2 \times S$ contingency table giving treatment $Z$ by outcome $r$ by stratum $s$. The test statistic is again the number of 1 responses among treated units, $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^T\mathbf{r} = \sum_{s=1}^{S} \sum_{i=1}^{n_s} Z_{si} r_{si}$. Under the null hypothesis, the contribution from stratum $s$, that is, $\sum_{i=1}^{n_s} Z_{si} r_{si}$, again has a hypergeometric distribution, and (2.5) is the distribution of the sum of $S$ independent hypergeometric variables. The Mantel–Haenszel statistic yields an approximation to the distribution of $\mathbf{Z}^T\mathbf{r}$ based on its expectation and variance, as described in more general terms in the next section. One technical attraction of this statistic is that the large sample approximation tends to work well for a $2 \times 2 \times S$ table with large $N$ even if $S$ is also large, so there may be few subjects in each of the $S$ tables. In particular, the statistic is widely used in matching with multiple controls, in which case $m_s = 1$ for each $s$.

*McNemar's* (1947) *test* is for paired binary data, that is, for $S$ pairs with $n_s = 2$, $m_s = 1$, and $r_{si} = 1$ or $r_{si} = 0$. The statistic is, yet again, the number of 1 responses among treated units; that is, $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^T\mathbf{r}$. McNemar's statistic is, in fact, a special case of the Mantel–Haenszel statistic, though the $2 \times 2 \times S$ table now describes $S$ pairs and certain simplifications are possible. In particular, the distribution of $\mathbf{Z}^T\mathbf{r}$ in (2.5) is that of a constant plus a certain binomial random variable.

Developing these methods for $2 \times 2 \times S$ tables in a different way, Birch (1964) and Cox (1966, 1970) show that these three tests with binary responses possess an optimality property, so there is a sense in which Fisher's exact test, the Mantel–Haenszel test, and McNemar's test are the best tests for the problems they address. Specifically, they show that the test statistic $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^T\mathbf{r}$ together with the significance level (2.5) is a uniformly most powerful unbiased test against alternatives defined in terms of constant odds ratios.

*Mantel's* (1963) *extension* of the Mantel–Haenszel test is for responses $r_{si}$ that are confined to a small number of values representing a numerical scoring of several ordered categories. As an example of such an outcome, the New York Heart Association classifies coronary patients into one of four categories based on the degree to which the patient is limited in physical activity by coronary symptoms such as chest pain. The categories are:

(1) no limitation of physical activity;

(2) slight limitation, comfortable at rest, but ordinary physical activity results in pain or other symptoms;

(3) marked limitation, minor activities result in coronary symptoms; and

(4) unable to carry on any physical activity without discomfort, which may be present even at rest.

The outcome $r_{si}$ for a patient is then one of the integers 1, 2, 3, or 4. In this case the data might be recorded as a $2 \times 4 \times S$ contingency table for $Z \times r \times s$. Mantel's test statistic is the sum of the response scores for treated units; that is, $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{r}$. Birch (1965) shows that the test is optimal in a certain sense.

In the case of a single stratum, $S = 1$, *Wilcoxon's* (1945) *rank sum test* is commonly used to compare outcomes taking many numerical values. In this test, the responses are ranked from smallest to largest. If all $N$ responses were different numbers, the ranks would be the numbers $1, 2, \ldots, N$. If some of the responses were equal, then the average of their ranks would be used. Write $q_i$ for the rank of $r_i$, and write $\mathbf{q} = (q_1, \ldots, q_N)^{\mathrm{T}}$. For instance, if $N = 4$, and $r_1 = 2.3$, $r_2 = 1.1$, $r_3 = 2.3$, and $r_4 = 7.9$, then $q_1 = 2.5$, $q_2 = 1$, $q_3 = 2.5$, and $q_4 = 4$, since $r_2$ is smallest, $r_4$ is largest, and $r_1$ and $r_3$ share the ranks 2 and 3 whose average rank is $2.5 = (2+3)/2$. Note that the ranks $\mathbf{q}$ are a function of the responses $\mathbf{r}$ which are fixed if the treatment has no effect, so $\mathbf{q}$ is also fixed. The rank sum statistic is the sum of the ranks of the treated observations, $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{q}$, and its significance level is determined from (2.5). The properties of the rank sum test have been extensively studied; for instance, see Lehmann (1975, §1) or Hettmansperger (1984, §3). Wilcoxon's rank sum test is equivalent to the *Mann and Whitney* (1947) *test*.

In the case of $S$ matched pairs with $n_s = 2$ and $m_s = 1$ for $s = 1, \ldots, S$, Wilcoxon's (1945) signed rank test is commonly used for responses taking many values. Here, $(Z_{s1}, Z_{s2}) = (1, 0)$ if the first unit in pair $s$ received the treatment or $(Z_{s1}, Z_{s2}) = (0, 1)$ if the second unit received the treatment. In this test, the absolute differences in responses within pairs $|r_{s1} - r_{s2}|$ are ranked from 1 to $S$, with average ranks used for ties. Let $d_s$ be the rank of $|r_{s1} - r_{s2}|$ thus obtained. The signed rank statistic is the sum of the ranks for pairs in which the treated unit had a higher response than the control unit.

To write this formally, let $c_{s1} = 1$ if $r_{s1} > r_{s2}$ and $c_{s1} = 0$ otherwise, and similarly, let $c_{s2} = 1$ if $r_{s2} > r_{s1}$ and $c_{s2} = 0$ otherwise, so $c_{s1} = c_{s2} = 0$ if $r_{s1} = r_{s2}$. Then $Z_{s1}c_{s1} + Z_{s2}c_{s2}$ equals 1 if the treated unit in pair $s$ had a higher response than the control unit, and equals zero otherwise. It follows that the signed rank statistic is $\sum_{s=1}^{S} d_s \sum_{i=1}^{2} c_{si} Z_{si}$. Note that $d_s$ and $c_{si}$ are functions of $\mathbf{r}$ and so are fixed under the null hypothesis of no treatment effect. Also, if $r_{s1} = r_{s2}$, then pair $s$ contributes zero to the value of the statistic no matter how treatments are assigned. As with the rank sum test, the signed rank test is widely used and has been extensively studied; for instance, see Lehmann (1975, §3) or Hettmansperger (1984, §2). Section 3.2.4 below contains a numerical example using the sign-rank statistic in an observational study.

For stratified responses, a method that is sometimes used involves calculating the rank sum statistic separately in each of the $S$ strata, and taking the sum of these $S$ rank sums as the test statistic. This is the *stratified rank sum statistic*. It is easily checked that this statistic has the form $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{q}$ resembling the rank sum statistic; however, the ranks in $\mathbf{q}$ are no longer a permutation of the numbers $1, 2, \dots, N$, but rather of the numbers $1, \dots, n_1, 1, \dots, n_2, \dots, 1, \dots, n_S$, with adjustments for ties if needed. Also $\Omega$ has changed.

Hodges and Lehmann (1962) find the stratified rank sum statistic to be inefficient when $S$ is large compared to $N$. In particular, for paired data with $S = N/2$, the stratified rank test is equivalent to the sign test, which in turn is substantially less efficient than the signed rank test for data from short-tailed distributions such as the Normal. They suggest as an alternative the method of aligned ranks: the mean in each stratum is subtracted from the responses in that stratum creating aligned responses that are ranked from 1 to $N$, momentarily ignoring the strata. Writing $\mathbf{q}$ for these aligned ranks, the *aligned rank statistic* is the sum of the aligned ranks in the treated group, $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{q}$. See also Lehmann (1975, §3.3).

Another statistic is the *median test*. Let $c_{si} = 1$ if $r_{si}$ is greater than the median of the responses in stratum $s$ and let $c_{si} = 0$ otherwise, and let $\mathbf{c}$ be the $N$-tuple containing the $c_{si}$. Then $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{c}$ is the number of treated responses that exceed their stratum medians. With a single stratum, $S = 1$, the median test is quite good in large samples if the responses have a double exponential distribution, a distribution with a thicker tail than the normal; see, for instance, Hettmansperger (1984, §3.4, p. 146) and the more critical comments by Freidlin and Gastwirth (2000). In this test, the median is sometimes replaced by other quantiles or other measures of location.

Start with any statistic $t(\mathbf{Z}, \mathbf{r})$ and the randomization distribution of $t(\mathbf{Z}, \mathbf{r})$ may be determined from (2.5). This is true even of statistics that are commonly referred to a theoretical distribution instead, for instance, the *two-sample* or *paired t-tests*, among others. Welch (1937) and Wilk (1955) studied the relationship between the randomization distribution and the theoretical distribution of statistics that were initially derived from assumptions of Normally and independently distributed responses. They suggest that the theoretical distribution may be viewed as a computationally convenient approximation to the desired but computationally difficult randomization distribution. That is, they suggest that $t$-tests, like rank tests or Mantel–Haenszel tests, may be justified solely on the basis of the use of randomization in the design of an experiment, without reference to Normal independent errors. These findings depend on the good behavior of moments of sums of squares of responses over the randomization distribution; therefore, they depend on the absence of extreme responses. Still, the results are important as a conceptual link between Normal theory and randomization inference.

### *2.4.4 Classes of Test Statistics*

The similarity among the commonly used test statistics in §2.4.3 is striking but not accidental. In this book, these statistics are not discussed individually, except when used in examples. The important properties of the methods are shared by large classes of statistics, so it is both simpler and less repetitive to discuss the classes.

Though invented by different people at different times for different purposes, the commonly used statistics in §2.4.3 are similar for good reason. As the sample size $N$ increases, the number $K$ of treatment assignments in $\Omega$ grows very rapidly, and the direct calculation in (2.5) becomes very difficult to perform, even with the fastest computers. To see why this is true, take the simplest case consisting of one stratum, $S = 1$, and an equal division of the $n$ subjects into $m = n/2$ treated subjects and $m = n/2$ controls. Then there are $K = \binom{n}{n/2}$ treatment assignments in $\Omega$. If one more unit is added to the experiment, increasing the sample size to $n+1$, then $K$ is increased by a factor of $(n+1) / \{(n/2)+1\}$, that is, $K$ nearly doubles. Roughly speaking, if the fastest computer can calculate (2.5) directly for at most a sample of size $n$, and if computing power doubles every year for 10 years, then 10 years hence computing power will be $2^{10} = 1024$ times greater than today and it will be possible to handle a sample of size $n+10$. Direct calculation of (2.5) is not practical for large $n$.

The usual solution to this problem is to approximate (2.5) using a large sample or asymptotic approximation. The most common approximations use the moments of the test statistic, its expectation and variance, and sometimes higher moments. The needed moments are easily derived for certain classes of statistics, including all those in §2.4.3.

As an alternative to asymptotic approximation, there are several proposals for computing (2.5) exactly, but they are not, as yet, commonly used. One is to compute (2.5) exactly but indirectly using clever computations that avoid working with the set $\Omega$. For some statistics this can be done by calculating the characteristic function of the test statistic and inverting it using the fast Fourier transform; see Pagano and Tritchler (1983). A second approach is to design experiments differently so that $\Omega$ is a much smaller set, perhaps containing 10,000 or 100,000 treatment assignments. In this case, direct calculation is possible and any test statistic may be used; see Tukey (1985) for discussion.

The first class of statistics will be called *sum statistics* and they are of the form $t(\mathbf{Z}, \mathbf{r}) = \mathbf{Z}^{\mathrm{T}}\mathbf{q}$, where $\mathbf{q}$ is some function of $\mathbf{r}$. A sum statistic sums the scores $q_{si}$ for treated units. All of the statistics in §2.4.4 are sum statistics for suitable choices of $\mathbf{q}$. In Fisher's exact test, the Mantel–Haenszel test, and McNemar's test, $\mathbf{q}$ is simply equal to $\mathbf{r}$. In the rank sum test, $\mathbf{q}$ contains the ranks of $\mathbf{r}$. In the median test, $\mathbf{q}$ is the vector of ones and zeros identifying responses $r_{si}$ that exceed stratum medians. In the signed rank statistic, $q_{si} = d_s c_{si}$.

Simple formulas exist for the moments of sum statistics under the null hypothesis that the treatment is without effect. In this case, $\mathbf{r}$ is fixed, so $\mathbf{q}$ is also fixed. The moment formulas use the properties of simple random sampling without replacement. Recall that a simple random sample without replacement of size $m$ from a population of size $n$ is a random subset of $m$ elements from a set with $n$ elements where each of the $\binom{n}{m}$ subsets of size $m$ has the same probability $1/\binom{n}{m}$. Cochran (1963) discusses simple random sampling. In a uniform randomized experiment, the $m_s$ treated units in stratum $s$ are a simple random sample without replacement from the $n_s$ units in stratum $s$. The following proposition is proved in Problem 4.

**Proposition 2** *In a uniform randomized experiment, if the treatment has no effect, the expectation and variance of a sum statistic* $\mathbf{Z}^{\mathrm{T}}\mathbf{q}$ *are*

$$E(\mathbf{Z}^{\mathrm{T}}\mathbf{q}) = \sum_{s=1}^{S} m_s \bar{q}_s,$$

*and*

$$\operatorname{var}(\mathbf{Z}^{\mathrm{T}}\mathbf{q}) = \sum_{s=1}^{S} \overbrace{\frac{m_s(n_s - m_s)}{n_s(n_s - 1)}}^{\bar{z}_s(1-\bar{z}_s)\frac{n_s}{n_s-1}} \sum_{i=1}^{n_s} (q_{si} - \bar{q}_s)^2,$$

*where*

$$\bar{q}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} q_{si}.$$

Moments are easily determined for sum statistics, but other classes of statistics have other useful properties. The first such class, the *sign-score statistics*, is a subset of the sum statistics. A statistic is a sign-score statistic if it is of the form $t(\mathbf{Z}, \mathbf{r}) = \sum_{s=1}^{S} d_s \sum_{i=1}^{n_s} c_{si} Z_{si}$, where $c_{si}$ is binary, $c_{si} = 1$ or $c_{si} = 0$, and both $d_s$ and $c_{si}$ are functions of $\mathbf{r}$. Fisher's exact test, the

Mantel–Haenszel test, and McNemar's test are sign-score statistics with $d_s = 1$ and $c_{si} = r_{si}$. The signed rank and median test statistics are also sign-score statistics, but with $c_{si}$ and $d_s$ defined differently. A sign-score statistic is a sum statistic with $q_{si} = d_s c_{si}$, but many sum statistics, including the rank sum statistic, are not sign-score statistics. In Chapter 4, certain calculations are simpler for sign-score statistics than for certain other sum statistics, and this motivates the distinction.

Another important class of statistics is the class of *arrangement increasing functions* of $\mathbf{Z}$ and $\mathbf{r}$, which are defined in a moment. Informally, a statistic $t(\mathbf{Z}, \mathbf{r})$ is arrangement-increasing if it increases in value as the coordinates of $\mathbf{Z}$ and $\mathbf{r}$ are rearranged into an increasingly similar order within each stratum. In fact, all of the statistics in §2.4.3 are arrangement-increasing, so anything that is true of arrangement-increasing statistics is true of all the commonly used statistics in §2.3.2. Hollander, Proschan, and Sethuraman (1977) discuss many properties of arrangement-increasing functions.

A few preliminary terms are useful. The numbers $S$ and $n_s$, $s = 1, \ldots, S$ with $N = \sum n_s$, are taken as given. A *stratified $N$-tuple* $\mathbf{a}$ is an $N$-tuple in which the $N$ coordinates are divided into $S$ strata with $n_s$ coordinates in stratum $s$, where $a_{si}$ is the $i$th of the $n_s$ coordinates in stratum $s$. For instance, $\mathbf{Z}$ and $\mathbf{r}$ are each stratified $N$-tuples. If $\mathbf{a}$ is a stratified $N$-tuple, and if $i$ and $j$ are different positive integers less than or equal to $n_s$, then let $\mathbf{a}_{sij}$ be the stratified $N$-tuple formed from $\mathbf{a}$ by interchanging $a_{si}$ and $a_{sj}$, that is, by placing the value $a_{sj}$ in the $i$th position in stratum $s$ and placing the value $a_{si}$ in the $j$th position in stratum $s$. To avoid repetition, whenever the symbol $\mathbf{a}_{sij}$ appears, it is assumed without explicit mention that the subscripts are appropriate, so $s$ is a positive integer between 1 and $S$ and $i$ and $j$ are different positive integers less than or equal to $n_s$. A function $f(\mathbf{a}, \mathbf{b})$ of two stratified $N$-tuples is *invariant* if $f(\mathbf{a}, \mathbf{b}) = f(\mathbf{a}_{sij}, \mathbf{b}_{sij})$ for all $s$, $i$, $j$, so renumbering units in the same stratum does not change the value of $f(\mathbf{a}, \mathbf{b})$. For instance, the function $\mathbf{z}^T \mathbf{q}$ is an invariant function of $\mathbf{z}$ and $\mathbf{q}$.

**Definition 3** *An invariant function $f(\mathbf{a},\mathbf{b})$ of two stratified $N$-tuples is arrangement-increasing (or AI) if $f(\mathbf{a}, \mathbf{b}_{sij}) \geq f(\mathbf{a},\mathbf{b})$ whenever*

$$(a_{si} - a_{sj}) \cdot (b_{si} - b_{sj}) \leq 0.$$

Notice what this definition says. Consider the $i$th and $j$th unit in stratum $s$. If $(a_{si} - a_{sj})(b_{si} - b_{sj}) < 0$, then of these two units, the one with the higher value of $a$ has the lower value of $b$, so these two coordinates are out of order. However, in $\mathbf{a}$ and $\mathbf{b}_{sij}$, these two coordinates are in the same order, for $b_{si}$ and $b_{sj}$ have been interchanged. The definition says that an arrangement increasing function will be larger, or at least no smaller, when these two coordinates are switched into the same order.

TABLE 2.3. A Hypothetical Example Showing an Arrangement-Increasing Statistic.

| $i$ | | $\mathbf{z}$ | $\mathbf{q}$ | $\mathbf{q}_{23}$ |
|---|---|---|---|---|
| 1 | Treated | 1 | 4 | 4 |
| 2 | Treated | 1 | 2 | 3 |
| 3 | Control | 0 | 3 | 2 |
| 4 | Control | 0 | 1 | 1 |
| Rank sum | | | 6 | 7 |

Notice also what the definition says when $(a_{si}-a_{sj})(b_{si}-b_{sj}) = 0$. In this case, either $a_{si} = a_{sj}$ or $b_{si} = b_{sj}$ or both. In this case, $f(\mathbf{a}, \mathbf{b}_{sij}) = f(\mathbf{a}, \mathbf{b})$.

Consider some examples. The function $\mathbf{z}^T\mathbf{q}$ is arrangement-increasing as a function of $\mathbf{z}$ and $\mathbf{q}$. To see this, note that $\mathbf{z}^T\mathbf{q}_{sij}-\mathbf{z}^T\mathbf{q} = (z_{si}q_{sj}+z_{sj}q_{si})-(z_{si}q_{si}+z_{sj}q_{sj}) = -(z_{si}-z_{sj})(q_{si}-q_{sj})$, so if $(z_{si}-z_{sj})(q_{si}-q_{sj}) \leq 0$ then $\mathbf{z}^T\mathbf{q}_{sij} - \mathbf{z}^T\mathbf{q} \geq 0$. This shows $\mathbf{z}^T\mathbf{q}$ is arrangement-increasing.

Table 2.3 is a small illustration for the rank sum statistic with a single stratum, $S = 1$, $n = 4$ units, of whom $m = 2$ received the treatment. Here, $(z_2 - z_3)(q_2 - q_3) = (1 - 0)(2 - 3) = -1 \leq 0$, and the rank sum $\mathbf{z}^T\mathbf{q} = 6$ is increased to $\mathbf{z}^T\mathbf{q}_{23} = 7$ by interchanging $q_2$ and $q_3$.

As a second example, consider the function $t(\mathbf{z}, \mathbf{r}) = \mathbf{z}^T\mathbf{q}$, where $\mathbf{q}$ is a function of $\mathbf{r}$, which may be written explicitly as $\mathbf{q}(\mathbf{r})$. Then $t(\mathbf{z}, \mathbf{r})$ may or may not be arrangement-increasing in $\mathbf{z}$ and $\mathbf{r}$ depending upon how $\mathbf{q}(\mathbf{r})$ varies with $\mathbf{r}$. The common statistics in §2.4.3 all have the following two properties:

(i) permute $\mathbf{r}$ within strata and $\mathbf{q}$ is permuted in the same way; and

(ii) within each stratum, larger $r_{si}$ receive larger $q_{si}$.

One readily checks that $t(\mathbf{z}, \mathbf{r}) = \mathbf{z}^T\mathbf{q}$ is arrangement-increasing if $\mathbf{q}(\mathbf{r})$ has these two properties, because the first property ensures that $t(\mathbf{z}, \mathbf{r})$ is invariant, and the second ensures that $r_{si} - r_{sj} \geq 0$ implies $q_{si} - q_{sj} \geq 0$, so $(z_{si} - z_{sj})(r_{si} - r_{sj}) \leq 0$ implies $(z_{si} - z_{sj})(q_{si} - q_{sj}) \leq 0$, and the argument of the previous paragraph applies. The important conclusion is that all of the statistics in §2.4.3 are arrangement-increasing.

In describing the behavior of a statistic when the null hypothesis does not hold and instead the treatment has an effect, a final class of statistics is useful. Many statistics that measure the size of the difference between treated and control groups would tend to increase in value if responses in the treated group were increased and those in the control group were decreased. Statistics with this property will be called effect increasing, and the idea will now be expressed this formally. A treated unit has $2Z_{si}-1 = 1$, since $Z_{si} = 1$, and a control unit has $2Z_{si}-1 = -1$ since $Z_{si} = 0$. Let $\mathbf{z} \in \Omega$

TABLE 2.4. Hypothetical Example of an Effect Increasing Statistic.

| $i$ | | $z_i$ | $2z_i - 1$ | $r_i$ | $r_i^*$ |
|---|---|---|---|---|---|
| 1 | Treated | 1 | 1 | 5 | 6 |
| 2 | Treated | 1 | 1 | 2 | 4 |
| 3 | Control | 0 | $-1$ | 3 | 2 |
| 4 | Control | 0 | $-1$ | 1 | 1 |
| Rank sum | | | | 6 | 7 |

be a possible treatment assignment and let **r** and **r**$^*$ be two possible values of the $N$-tuple of responses such that $(r_{si}^* - r_{si})(2z_{si} - 1) \geq 0$ for all $s$, $i$. With treatments given by **z**, this says that $r_{si}^* \geq r_{si}$ for every treated unit and $r_{si}^* \leq r_{si}$ for every control unit. In words, if higher responses indicated favorable outcomes, then every treated unit does better with **r**$^*$ than with **r**, and every control does worse with **r**$^*$ than with **r**. That is, the difference between treated and control groups looks larger with **r**$^*$ than with **r**. The test statistic is *effect increasing* if $t(\mathbf{z}, \mathbf{r}) \leq t(\mathbf{z}, \mathbf{r}^*)$ whenever **r** and **r**$^*$ are two possible values of the response such that $(r_{si}^* - r_{si})(2z_{si} - 1) \geq 0$ for all $s$, $i$. All of the commonly used statistics in §2.4.3 are effect increasing.

Table 2.4 contains a small hypothetical example to illustrate the idea of an effect increasing statistic. Here there is a single stratum, $S = 1$, and four subjects, $n = 4$, of whom $m = 2$ received the treatment. Notice that when $r_i$ and $r_i^*$ are compared, treated subjects have $r_i^* \geq r_i$ while controls have $r_i^* \leq r_i$. If the responses are ranked 1, 2, 3, 4, and the ranks in the treated group are summed to give Wilcoxon's rank sum statistic, then the rank sum is larger for $r_i^*$ than for $r_i$.

In summary, this section has considered four classes of statistics:

(i) the sum statistics;

(ii) the arrangement-increasing statistics;

(iii) the effect increasing statistics; and

(iv) the sign-score statistics.

All of the commonly used statistics in §2.4.3 are members of the first three classes, and most are sign-score statistics; however, the rank sum statistic, the stratified rank sum statistic, and Mantel's extension are not sign-score statistics.

### 2.4.5 *No Effect Means No Effect

No effect means no effect. A nonzero effect that varies from one unit to the next and that is hard to fathom or predict is, nonetheless, a nonzero effect. It may not be an immediately useful effect, but it is an effect, perhaps an effect that can someday be understood, tamed, and made useful.

Empirically, it may be difficult to discern erratic unsystematic effects, but logically they are distinct from no effect.

To emphasize this point, consider the extreme case. Suppose that we somehow discerned that the treatment erratically benefits some patients and harms others, but that we have no way of predicting who will benefit or who will be harmed, so the average effect of the treatment is essentially zero in every large group of patients defined by pretreatment variables. In point of fact, it is very difficult to discern something like this, unless we covertly introduce more information that does distinguish these supposedly indistinguishable patients. Suppose, however, we can discern this, perhaps because the treatment produces one of two easily distinguished biochemical reactions, one beneficial, the other harmful, and neither reaction is ever seen among controls; however, we are completely at a loss to identify in advance those patients who will have beneficial reactions. This is a nonzero treatment effect, perhaps not a very useful one given current knowledge, but a nonzero effect nonetheless. What would a scientist do with such an effect? Might the scientist sometimes return with the treatment to the laboratory in an effort to understand why only some patients exhibit the beneficial biochemical reaction? In contrast, no treatment effect—really no treatment effect—would send the scientist in search of another treatment.

No effect is one hypothesis among many. It is rarely, perhaps never, sufficient to know whether the null hypothesis of no treatment effect is compatible with observed data. And yet, it is typically of interest to know this along with much more. Section 2.5 and Chapter 5 discuss models for treatment effects and associated methods of inference, including confidence intervals.

Fisher (1935) and Neyman (1935), two brilliant founders of statistics, did not agree about the meaning of the null hypothesis of no treatment effect. The hypothesis of no effect as I have described it is Fisher's version. Fisher's conception is particular: randomization justifies causal inferences about particular treatment effects, on particular units, at a particular time, under particular circumstances. Change the units or the times or the circumstances and the findings may change to an extent not adequately addressed by statistical standard errors. These standard errors measure one very important source of uncertainty, namely, uncertainty about how units would have responded to a treatment they did not receive, that is, uncertainty about the effects caused by the treatment. Campbell and Stanley (1963) say that randomization ensures *internal validity* but not *external validity*; see §2.7.1 and the discussion of efficacy and effectiveness in §5.4. Neyman's (1935, p. 110) conception is general: we can "repeat the experiment indefinitely without any change of vegetative conditions or of arrangement so that ... the yields from this plot will form a population ...." For Neyman, the variations we do not understand become, by assumption, variations from sampling a population. In point of fact, we cannot repeat the experiment indefinitely, and we cannot ensure the same experimental conditions, but this conception concerns a hypothetical world in which we can. This was not a disagreement about matters of fact, but about matters of art, the art of developing statistical concepts for scientific applications.

In most cases, their disagreement is entirely without technical consequence: the same procedures are used, and the same conclusions are reached. Perhaps this is expressed most beautifully by Lehmann (1959, §5). First, Lehmann (1959, §5.7, Theorem 3) shows that inferences under a population model can be distribution-free only if they are made particular by conditioning on observed responses, yielding Fisher's randomization test. Lehmann (1959, §5.8) then uses a population model and the Neyman—Pearson lemma to obtain most powerful permutation tests; that is, he uses Neyman's conception to obtain the best tests of the type Fisher was proposing. Whatever Fisher and Neyman may have thought, in Lehmann's text they work together. The importance to mathematical statistics and to science of infinite population models and Neyman's contributions are, today, surely unquestioned.

And yet, when one is thinking about the science of an experiment, it is surely true that random assignment of treatments justifies inferences that are particular, that is, particular to certain units at certain times under certain circumstances. If the inference reaches beyond that to infinite populations extending into the indefinite future, then this has been accomplished by assuming those populations into existence, and assuming away much that is true of the world we actually inhabit. In those instances where their conceptions point in scientifically different directions—for instance, the unpredictable but distinguishable biochemical reactions above—it seems to me that Fisher's conception more closely describes how scientists think and work. Much that we cannot currently predict and do not currently fathom is not random error. The variation we do not fathom today we intend to decipher tomorrow.

## 2.5 Simple Models for Treatment Effects

### *2.5.1 Responses When the Treatment Has an Effect*

If the treatment has an effect, then the observed $N$-tuple of responses for the $N$ units will be different for different treatment assignments $\mathbf{z} \in \Omega$—this is what it means to say the treatment has an effect. In earlier sections, the null hypothesis of no treatment effect was assumed to hold, so the observed responses were fixed, not varying with $\mathbf{z}$, and the response was written $\mathbf{r}$. When the null hypothesis of no effect is not assumed to hold, the response changes with $\mathbf{z}$, and the response observed when the treatment assignment is $\mathbf{z} \in \Omega$ will be written $\mathbf{r}_\mathbf{z}$. The null hypothesis of no treatment effect says that $\mathbf{r}_\mathbf{z}$ does not vary with $\mathbf{z}$, and instead $\mathbf{r}_\mathbf{z}$ is a constant the same for all $\mathbf{z}$; in this case, $\mathbf{r}$ was written for this constant. Notice that, for each

$\mathbf{z} \in \Omega$, the response $\mathbf{r_z}$ is some nonrandom $N$-tuple—probability has not yet entered the discussion. Write $r_{si\mathbf{z}}$ for the $(s, i)$ coordinate of $\mathbf{r_z}$, that is, for the response of the $i$th unit in stratum $s$ when the $N$ units receive the treatment assignment $\mathbf{z}$.

To make this definite, return for a moment to Fisher's Lady tasting tea. If the Lady could not discriminate at all, then no matter how milk is added to the cup—that is, no matter what $\mathbf{z}$ is—she will classify the cups in the same way; that is, she will give the same binary 8-tuple of responses $\mathbf{r}$. On the other hand, if she discriminates perfectly, always classifying cups correctly, then her 8-tuple of responses will vary with $\mathbf{z}$; indeed, the responses will match the treatment assignments so that $\mathbf{r_z} = \mathbf{z}$.

If treatments are randomly assigned, then the treatment assignment $\mathbf{Z}$ is a random variable, so the observed responses are also random variables as they depend on $\mathbf{Z}$. Specifically, the observed response is the random variable $\mathbf{r_Z}$, that is, one of the many possible $\mathbf{r_z}$, $\mathbf{z} \in \Omega$, selected by picking a treatment assignment $\mathbf{Z}$ by the random mechanism that governs the experiment. Write $\mathbf{R} = \mathbf{r_Z}$ for the observed response, where $\mathbf{R}$ like $\mathbf{Z}$ is a random variable.

In principle, each possible treatment assignment $\mathbf{z} \in \Omega$ might yield a pattern of responses $\mathbf{r_z}$ that is unrelated to the pattern observed with another $\mathbf{z}$. For instance, in a completely randomized experiment with 50 subjects divided into two groups of 25, there might be $|\Omega| = \binom{50}{25} \doteq 1.3 \times 10^{14}$ different and unrelated 50-tuples $\mathbf{r_z}$. Since it is difficult to comprehend a treatment effect in such terms, we look for regularities, patterns, or models of the behavior of $\mathbf{r_z}$ as $\mathbf{z}$ varies over $\Omega$. The remainder of §2.5 discusses the most basic models for $\mathbf{r_z}$ as $\mathbf{z}$ varies over $\Omega$. Chapter 5 discusses additional models for treatment effects.

### *2.5.2 No Interference Between Units*

A first model is that of "no interference between units" which means that "the observation on one unit should be unaffected by the particular assignment of treatments to the other units" (Cox, 1958a, §2.4). Rubin (1986) calls this SUTVA for the "stable unit treatment value assumption." Formally, no interference means that $r_{si\mathbf{z}}$ varies with $z_{si}$ but not with the other coordinates of $\mathbf{z}$. In other words, the response of the $i$th unit in stratum $s$ depends on the treatment assigned to this unit, but not on the treatments assigned to other units, so this unit has only two possible values of the response rather than $|\Omega|$ possible values. When this model is assumed, write $r_{Tsi}$ and $r_{Csi}$ for the responses of the $i$th unit in stratum $s$ when assigned, respectively, to treatment or control; that is, $r_{Tsi}$ is the common value of $\mathbf{r}_{si\mathbf{z}}$ for all $\mathbf{z} \in \Omega$ with $z_{si} = 1$, and $r_{Csi}$ is the common value of $r_{si\mathbf{z}}$ for all $z \in \Omega$ with $z_{si} = 0$. Then the observed response from the $i$th unit in stratum $s$ is $R_{si} = r_{Tsi}$ if $Z_{si} = 1$ or $R_{si} = r_{Csi}$ if $Z_{si} = 0$, which may also be written $R_{si} = Z_{si}r_{Tsi} + (1 - Z_{si})r_{Csi}$. This model, with one potential response for each unit under each treatment, has been important both to experimental design—see Neyman (1923), Welch (1937), Wilk (1955), Cox (1958b, §5), and Robinson (1973)—and to causal inference more generally—see Rubin (1974, 1977) and Holland (1986). When there is no interference between units, write $\mathbf{r}_T$ for $(r_{T11}, \ldots, r_{TS,n_S})^T$ and $\mathbf{r}_C$ for $(r_{C11}, \ldots, r_{CS,n_S})^T$.

"No interference between units" is a model and it can be false. No interference is often plausible when the units are different people and the treatment is a medical intervention with a biological response. In this case, no interference means that a medical treatment given to one patient affects only that patient, not other patients. That is often true. However, a vaccine given to many people may protect unvaccinated individuals by reducing the spread of a virus (so called herd immunity) and this is a form of interference. No interference is less plausible in some social settings, such a workplace or a classroom, where a reward given to one person may be visible to others, and may affect their behavior. No interference is often implausible when the strata are people and the units are repeated measures on a person; then a treatment given at one time may affect responses at later times; see Problem 2. In randomized single subject experiments, such as the Lady tasting tea, no interference is typically implausible.

### *2.5.3 The Model of an Additive Effect, and Related Models*

The model of an additive treatment effect assumes units do not interfere with each other, and the administration of the treatment raises the response of a unit by a constant amount $\tau$, so that $r_{Tsi} = r_{Csi} + \tau$ for each $s$, $i$. The principal attraction of the model is that there is a definite parameter to estimate, namely, the additive treatment effect $\tau$. As seen in §2.7, in a uniform randomized experiment, many estimators do indeed estimate $\tau$ when this model holds.

In understanding the model of an additive treatment effect, it is important to keep in mind that the pair of responses, $(r_{Tsi}, r_{Csi})$, is never jointly observed for one unit $(s, i)$. Therefore the model of an additive effect, $r_{Tsi} = r_{Csi} + \tau$, cannot be checked directly by comparing $r_{Tsi}$ and $r_{Csi}$ for particular units. The treatment $Z_{si}$ and the observed response $R_{si} = Z_{si}r_{Tsi} + (1 - Z_{si})r_{Csi}$ are observed, and one can check what the model, $r_{Tsi} = r_{Csi} + \tau$, implies about these observable quantities. In a completely randomized experiment with a single stratum, $S = 1$, dropping the $s$, the model of an additive treatment effect, $r_{Ti} = r_{Ci} + \tau$, implies that, as sample sizes $m$ and $n - m$ increase, the distribution of observed responses $R_i$ for treated units $Z_i = 1$ will be shifted by $\tau$ when compared to the distribution of observed responses $R_i$ for controls $Z_i = 0$, so the distributions will have the same shape and dispersion. That is, the histograms or boxplots would look the same, but one would be moved left or right relative to the other. This is a shift model, commonly used in nonparametrics; see

Lehmann (1975). In a uniform randomized experiment with several strata $S > 1$ and $r_{Tsi} = r_{Csi} + \tau$, the distribution of responses may have different shapes and dispersions in different strata, but within each stratum, the treated and control distributions are shifted by $\tau$. This is a fairly weak form of no interaction between treatment group and stratum in the $2 \times S$ table of observable distributions, and it implies much less about observable distributions than the analogous nonparametric analysis of variance model, which typically assumes a common shape and dispersion in all $2S$ cells. If the only data are $(Z_{si}, R_{si})$, does the additive model have content beyond its implications for observable distributions? See Problem 7.

Under the additive model, the observed response from the $i$th unit in stratum $s$ is $R_{si} = r_{Csi} + \tau Z_{si}$, or $\mathbf{R} = \mathbf{r}_C + \tau \mathbf{Z}$. It follows that the *adjusted responses*, $\mathbf{R} - \tau \mathbf{Z} = \mathbf{r}_C$, are fixed, not varying with the treatment assignment, $\mathbf{Z}$, so the adjusted responses satisfy the null hypothesis of no effect. This fact will be useful in drawing inferences about $\tau$.

There are many similar models, including the model of a multiplicative effect, $r_{Tsi} = \sigma r_{Csi}$. Chapter 5 discusses quite different models for treatment effects.

### *2.5.4 *Positive Effects and Larger Effects*

The model of an additive effect assumes a great deal about the relationship between $r_{Tsi}$ and $r_{Csi}$. At times, it is desirable to describe the behavior of statistical procedures while assuming much less. When there is no interference between units, an effect is a pair $(\mathbf{r}_T, \mathbf{r}_C)$ giving the responses of each unit under each treatment. Two useful concepts are positive effects and larger effects. Unlike the model of an additive treatment effect, positive effects and larger effects are meaningful not just for continuous responses, but also for binary responses, for ordinal responses, and as seen later in §2.8, for censored responses and multivariate responses.

A treatment has a *positive effect* if $r_{Tsi} \geq r_{Csi}$ for all units $(s, i)$ with strict inequality for at least one unit. A more compact way of writing this is that $(\mathbf{r}_T, \mathbf{r}_C)$ is a positive effect if $\mathbf{r}_T \geq \mathbf{r}_C$ with $\mathbf{r}_T \neq \mathbf{r}_C$. This says that application of the treatment never decreases a unit's response and sometimes increases it. For instance, there is a positive effect if the effect is additive and $\tau > 0$. Hamilton (1979) discusses this model in detail when the outcome is binary.

Consider two possible effects, say $(\mathbf{r}_T, \mathbf{r}_C)$ and $(\mathbf{r}_T^*, \mathbf{r}_C^*)$. Then $(\mathbf{r}_T^*, \mathbf{r}_C^*)$ is a *larger effect* than $(\mathbf{r}_T, \mathbf{r}_C)$ if $r_{Tsi}^* \geq r_{Tsi}$ and $r_{Csi}^* \leq r_{Csi}$ for all $s$, $i$. For instance, the simplest example occurs when the treatment effect is additive with the same responses under control, namely, $\mathbf{r}_C^* = \mathbf{r}_C$, $\mathbf{r}_T = \mathbf{r}_C + \tau \mathbf{1}$, and $\mathbf{r}_T^* = \mathbf{r}_C + \tau^* \mathbf{1}$, for in this case $(\mathbf{r}_T^*, \mathbf{r}_C^*)$ exhibits a *larger effect* than $(\mathbf{r}_T, \mathbf{r}_C)$ if $\tau^* \geq \tau$. In general, write $\mathbf{R}$ and $\mathbf{R}^*$ for the observed responses from, respectively, the effects $(\mathbf{r}_T, \mathbf{r}_C)$ and $(\mathbf{r}_T^*, \mathbf{r}_C^*)$, so $R_{si}^* = r_{Tsi}^*$ if $Z_{si} = 1$ and $R_{si}^* = r_{Csi}^*$ if $Z_{si} = 0$.

If a statistical test rejects the null hypothesis 5% of the time when it is true, one would hope that it would reject at least 5% of the time when it is false in the anticipated direction. Recall that a statistical test is *unbiased* against a collection of alternative hypotheses if the test is at least as likely to reject the null hypothesis when one of the alternatives is true as when the null hypothesis is true. The next proposition says that all of the common tests in §2.4.3 are unbiased tests against positive treatment effects, and the test statistic is larger when the effect is larger. The proposition is proved in somewhat more general terms in the appendix, §2.9.

**Proposition 4** *In a randomized experiment, a test statistic that is effect increasing yields an unbiased test of no effect against the alternative of a positive effect, and if* $(\mathbf{r}_T^*, \mathbf{r}_C^*)$ *is a larger effect than* $(\mathbf{r}_T, \mathbf{r}_C)$*, then* $t(\mathbf{Z}, \mathbf{R}^*) \geq t(\mathbf{Z}, \mathbf{R})$.

## 2.6 Confidence Intervals

### *2.6.1 Testing General Hypotheses*

So far, the test statistic $t(\mathbf{Z}, \mathbf{R})$ has been used to test the null hypothesis of no treatment effect. There is an extension to test hypotheses that specify a particular treatment effect. In §2.6.2, this extension is used to construct confidence intervals. As always, the confidence interval is the set of hypotheses not rejected by a test.

Consider testing the hypothesis $H_0 : \tau = \tau_0$ in the model of an additive effect, $\mathbf{R} = \mathbf{r}_C + \tau\mathbf{Z}$. The idea is as follows. If the null hypothesis $H_0 : \tau = \tau_0$ were true, then $\mathbf{r}_C = \mathbf{R} - \tau_0\mathbf{Z}$, so testing $H_0 : \tau = \tau_0$ is the same as testing that $\mathbf{R} - \tau_0\mathbf{Z}$ satisfies the null hypothesis of no treatment effect.

More precisely, if $\mathbf{r}_C$ were known, the probability, say $\alpha$, that $t(\mathbf{Z}, \mathbf{r}_C)$ is greater than or equal to some fixed number $T$ could be determined from (2.3). If the null hypothesis were true, then $\mathbf{r}_C$ would equal the *adjusted responses*, $\mathbf{R} - \tau_0\mathbf{Z}$, so under the null hypothesis, $\mathbf{r}_C$ can be calculated from $\tau_0$ and the observed data. If the hypothesis $H_0 : \tau = \tau_0$ is true, then the chance that $t(\mathbf{Z}, \mathbf{R} - \tau_0\mathbf{Z}) \geq T$ is $\alpha$, where $\alpha$ is calculated as described above with $\mathbf{r}_C = \mathbf{R} - \tau_0\mathbf{Z}$.

Now, suppose the null hypothesis is not true, say instead $\tau > \tau_0$, and consider the behavior of the above test. In this case, $\mathbf{R} = \mathbf{r}_C + \tau\mathbf{Z}$ and the adjusted responses $\mathbf{R} - \tau_0\mathbf{Z}$ equal $\mathbf{r}_C + (\tau - \tau_0)\mathbf{Z}$, so the adjusted responses will vary with the assigned treatment $\mathbf{Z}$. If a unit receives the treatment, it will have an adjusted response that is $\tau - \tau_0$ higher than if this unit receives the control. If the test statistic is effect increasing, as is true of all the statistics in §2.4.3, then $t(\mathbf{Z}, \mathbf{R} - \tau_0\mathbf{Z}) = t\{\mathbf{Z}, \mathbf{r}_C + (\tau - \tau_0)\mathbf{Z}\} \geq t(\mathbf{Z}, \mathbf{r}_C) = t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$, where the inequality follows from the definition of an effect increasing statistic. In words, if the null hypothesis is false and

TABLE 2.5. Example of Confidence Interval Computations.

| Unit | Control Response | Group | Observed Response | Adjusted Response | Ranks of Adjusted Responses |
|---|---|---|---|---|---|
| $i$ | $r_{Ci}$ | $Z_i$ | $R_i = r_{Ci} + \tau Z_i$ | $R_i - \tau_0 Z_i$ | $q_i$ |
| 1 | 2 | 1 | 9 | 8 | 7 |
| 2 | 1 | 0 | 1 | 1 | 1 |
| 3 | 3 | 0 | 3 | 3 | 2 |
| 4 | 4 | 0 | 4 | 4 | 3 |
| 5 | 0 | 1 | 7 | 6 | 5 |
| 6 | 4 | 1 | 11 | 10 | 8 |
| 7 | 1 | 1 | 8 | 7 | 6 |
| 8 | 5 | 0 | 5 | 5 | 4 |
| | | | $\tau = 7, \tau_0 = 1$ | | |

instead $\tau > \tau_0$, then an effect increasing test statistic $t(\mathbf{Z}, \mathbf{R} - \tau_0\mathbf{Z})$ will be larger with the incorrect $\tau_0$ than it would have been had we tested the correct value $\tau$.

Table 2.5 illustrates these computations with a rank sum test. It is a hypothetical uniform randomized experiment with $N = 8$ units, all in a single stratum $S = 1$, with $m = 4$ units assigned to treatment, and an additive treatment effect $\tau = 7$, though the null hypothesis incorrectly says $H_0 : \tau = \tau_0 = 1$. The rank sum computed from the adjusted responses $\mathbf{R} - 1\mathbf{Z}$ is $7 + 5 + 8 + 6 = 26$, which is the largest possible rank sum for $N = 8$, $m = 4$, and the one-sided significance level is $\binom{8}{4}^{-1} = 1/70 = 0.014$. The two-sided significance level is $2 \times 0.014 = 0.028$. After removing the hypothesized $\tau_0 = 1$ from treated units, the treated units continue to have higher responses than the controls.

## *2.6.2 Confidence Intervals by Inverting a Test*

Under the model of an additive treatment effect, $\mathbf{R} = \mathbf{r}_C + \tau\mathbf{Z}$, a $1 - \alpha$ confidence set for $\tau$ is obtaining by testing each value of $\tau$ as in §2.6.1 and collecting all values not rejected at level $\alpha$ into a set $A$. More precisely, $A$ is the set of values of $\tau$ that, when tested, yield significance levels or $P$-values greater than or equal to $\alpha$. For instance, in the example in Table 2.5, the value $\tau = 1$ would not be contained in a 95% confidence set. When the true value $\tau$ is tested, it is rejected with probability no greater than $\alpha$, so the random set $A$ contains the true $\tau$ with probability at least $1 - \alpha$. This is called "inverting" a test, and it is the standard way of obtaining a confidence set from a test; see, for instance, Cox and Hinkley (1974, §7.2) or Lehmann (1959, §3.5). For many test statistics, a two-sided test yields a confidence set that is an interval, whose endpoints may be determined by a line search, as illustrated in §4.3.5. Section 3.2.4 uses this confidence interval in an observational study of lead in the blood of children.

## 2.7 Point Estimates

### 2.7.1 Unbiased Estimates of the Average Effect

The most quoted fact about randomized experiments is that they lead to unbiased estimates of the average treatment effect. Take the simplest case, a uniform randomized experiment with a single stratum, with no interference between units. In this case, there are $m$ treated units, $N-m$ control units, $E(Z_i) = m/N$, $R_i = r_{Ti}$ if $Z_i = 1$, and $R_i = r_{Ci}$ if $Z_i = 0$. The difference between the mean response in the treated group, namely, $(1/m)\sum Z_i R_i$, and the mean response in the control group, namely, $\{1/(N-m)\}\sum(1-Z_i)R_i$, has expectation

$$E\left\{\sum \frac{Z_i R_i}{m} - \frac{(1-Z_i)R_i}{N-m}\right\} = E\left\{\sum \frac{Z_i r_{Ti}}{m} - \frac{(1-Z_i)r_{Ci}}{N-m}\right\}$$

$$= \sum \frac{(m/N)r_{Ti}}{m} - \frac{(1-m/N)r_{Ci}}{N-m} = \frac{1}{N}\sum r_{Ti} - r_{Ci},$$

and the last term is the average of the $N$ treatment effects $r_{Ti} - r_{Ci}$ for the $N$ experimental units. In words, the difference in sample means is unbiased for the average effect of the treatment. Notice carefully that this is true assuming only that there is no interference between units. There is no assumption that the treatment effect $r_{Ti} - r_{Ci}$ is constant from unit to unit, no assumption about interactions.

The estimate is unbiased for the average effect on the $N$ units in this study, namely, $(1/N)\sum r_{Ti} - r_{Ci}$, but this says nothing about the effect on other units not in the study. Campbell and Stanley (1963) say that a randomized experiment has *internal validity* in permitting inferences about effects for the $N$ units in the study, but it need not have *external validity* in that there is no guarantee that the treatment will be equally effective for other units outside the study; see also §2.4.5. The related issue of efficacy and effectiveness is discussed in §5.4.

The difference in sample means may be biased when there are two or more strata and the experimenter assigns disproportionately more subjects to the treatment in some strata than in others. However, there is an unbiased estimate that corrects the imbalance. It consists of calculating, within stratum $s$, the difference between the average response in the treated group, namely, $(1/m_s)\sum_i Z_{si}R_{si}$, and the average response in the control group, namely, $\{1/(n_s - m_s)\} \sum_i (1 - Z_{si}) R_{si}$, and weighting this difference by the proportion of units in stratum $s$, namely, $n_s/N$. The estimate, called *direct adjustment*, is then:

$$\sum_{s=1}^{S} \frac{n_s}{N} \sum_{i=1}^{n_s} \left\{ \frac{Z_{si} R_{si}}{m_s} - \frac{(1 - Z_{si}) R_{si}}{n_s - m_s} \right\}. \tag{2.6}$$

To check that (2.6) is unbiased, recall that, in a uniform randomized experiment, $Z_{si}$ has expectation $m_s/n_s$. It follows that (2.6) has expectation

$$\begin{aligned} E&\left[ \sum_{s=1}^{S} \frac{n_s}{N} \sum_{i=1}^{n_s} \left\{ \frac{Z_{si} R_{si}}{m_s} - \frac{(1 - Z_{si}) R_{si}}{n_s - m_s} \right\} \right] \\ &= E\left[ \sum_{s=1}^{S} \frac{n_s}{N} \sum_{i=1}^{n_s} \left\{ \frac{Z_{si} r_{Tsi}}{m_s} - \frac{(1 - Z_{si}) r_{Csi}}{n_s - m_s} \right\} \right] \\ &= \sum_{s=1}^{S} \frac{n_s}{N} \sum_{i=1}^{n_s} \left\{ \frac{(m_s/n_s) r_{Tsi}}{m_s} - \frac{(1 - m_s/n_s) r_{Csi}}{n_s - m_s} \right\} \\ &= \frac{1}{N} \sum r_{Ti} - r_{Ci}, \end{aligned}$$

so direct adjustment does indeed give an unbiased estimate of the average effect. In a very clear discussion, Rubin (1977) does calculations of this kind.

In effect, direct adjustment views the treated units and the control units as two stratified random samples from the $N$ units in the experiment. Then (2.6) is the usual stratified estimate of mean response to treatment in the population of $N$ units minus the usual estimate of the mean response to control in the population of $N$ units. Notice again that direct adjustment is unbiased for the average treatment effect even if that effect varies from unit to unit or from stratum to stratum. On the other hand, the average effect is but a summary of the effects, and not a complete description, when the effect varies from one stratum to another.

### *2.7.2 Hodges–Lehmann Estimates of an Additive Effect*

Under the model of an additive effect, $\mathbf{R} = \mathbf{r}_C + \tau\mathbf{Z}$, there are many estimates of $\tau$. One due to Hodges and Lehmann (1963) is closely tied to the test in §2.4 and the confidence interval in §2.6. Recall that $H_0 : \tau = \tau_0$ is tested using $t(\mathbf{Z}, \mathbf{R} - \tau_0\mathbf{Z})$, that is, by subtracting the hypothesized treatment effect $\tau_0\mathbf{Z}$ from the observed responses $\mathbf{R}$, and asking whether the adjusted responses $\mathbf{R} - \tau_0\mathbf{Z}$ appear to be free of a treatment effect. The Hodges–Lehmann estimate of $\tau$ is that value $\hat{\tau}$ such that the adjusted responses $\mathbf{R} - \hat{\tau}\mathbf{Z}$ appear to be exactly free of a treatment effect. Consider this in detail. Throughout this section, the experiment is a uniform randomized experiment.

Suppose that we can determine the expectation, say $\bar{\bar{t}}$, of the statistic $t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$ when calculated using the correct $\tau$, that is, when calculated from responses $\mathbf{R} - \tau\mathbf{Z}$ that have been adjusted so they are free of a treatment effect. For instance, in an experiment with a single stratum, the rank sum statistic has expectation $\bar{\bar{t}} = m(N+1)/2$ if the treatment has no effect. This is true because, in the absence of a treatment effect, the rank sum statistic is the sum of $m$ scores randomly selected from $N$ scores whose mean is $(N+1)/2$. In the same way, in a stratified experiment, the stratified rank sum statistic has expectation $\bar{\bar{t}} = \frac{1}{2}\sum m_s(n_s+1)$ in the absence of a treatment effect. In an experiment comprised of $S$ pairs, in the absence of a treatment effect, the expectation of the signed rank statistic is $\bar{\bar{t}} = (S+1)/4$, since we expect to sum half of $S$ scores which average $(S+1)/2$. In the absence of an effect, in an experiment with a single stratum, the difference in sample means (2.2) has expectation $\bar{\bar{t}} = 0$. In each of these cases, $\bar{\bar{t}}$ may be determined without knowing $\tau$, so there is a Hodges–Lehmann estimate.

Roughly speaking, the Hodges–Lehmann estimate is the solution $\hat{\tau}$ of the equation $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$. In other words, $\hat{\tau}$ is the value such that the adjusted responses $\mathbf{R} - \hat{\tau}\mathbf{Z}$ appear to be entirely free of a treatment effect, in the sense that the test statistic $t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ exactly equals its expectation in the absence of an effect.

If $t(\cdot,\cdot)$ is an effect increasing statistic, as is true of all of the statistics in §2.3, then $t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ is monotone decreasing as a function of $\hat{\tau}$ with $\mathbf{Z}$ and $\mathbf{R}$ fixed. This says: The larger the treatment effect $\hat{\tau}\mathbf{Z}$ removed from the observed responses $\mathbf{R}$, the smaller the statistic becomes. This is useful in solving $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$. If a $\hat{\tau}$ has been tried such that $\bar{\bar{t}} < t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$, then a larger $\hat{\tau}$ will tend to make $t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ smaller, moving it toward $\bar{\bar{t}}$. Similarly, if $\bar{\bar{t}} > t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$, then a smaller $\hat{\tau}$ is needed.

Problems arise immediately. For rank statistics, such as the rank sum and the signed rank, $t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ varies in discrete jumps as $\hat{\tau}$ is varied, so there may be no value $\hat{\tau}$ such that $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$. To see this, take a trivial case, a uniform experiment in one stratum, sample size $N = 2$, one treated unit $m = 1$. Then the rank sum statistic is either 1 or 2 depending upon which of the two units receive the treatment, but $\bar{\bar{t}} = 1.5$, so it is not possible to find a $\hat{\tau}$ such that $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$.

Not only may $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ have no solution $\hat{\tau}$, but it may have infinitely many solutions. If $t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ varies in discrete jumps, it will be constant for intervals of values of $\hat{\tau}$.

Hodges and Lehmann resolve these problems in the following way. They define the solution of an equation $\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$ as SOLVE$\{\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} -$

TABLE 2.6. Computing a Hodges–Lehmann Estimate.

| $\tau$ | 4.9999 | 5 | 5.0001 | 5.9999 | 6 | 6.0001 |
|---|---|---|---|---|---|---|
| $t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$ | 20 | 19 | 18 | 18 | 17 | 15 |

$\hat{\tau}\mathbf{Z})\}$ defined by

$$\begin{aligned}\hat{\tau} &= \text{SOLVE}\{\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})\} \\ &= \frac{\inf\{\tau : \bar{\bar{t}} > t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})\} + \sup\{\tau : \bar{\bar{t}} < t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})\}}{2}.\end{aligned}$$

This defines the Hodges–Lehmann estimate. Roughly speaking, if there is no exact solution, then average the smallest $\tau$ that is too large and the largest $\tau$ that is too small.

Consider the small example in Table 2.5. Under the null hypothesis of no effect, the rank sum statistic has expectation $\bar{\bar{t}} = m(N+1)/2 = 4(8+1)/2 = 18$, that is, half of the sum of all eight ranks, $36 = 1 + 2 + \cdots + 8$. Table 2.6 gives values of $t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$ for several values of $\tau$. As noted, since $t(\cdot, \cdot)$ is effect increasing, in Table 2.6, $t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})$ decreases in $\tau$. We want as our estimate a value $\hat{\tau}$ such that $18 = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})$, but the table indicates that any value between 5 and 6 will do. As the table suggests, $\inf\{\tau : \bar{\bar{t}} > t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})\} = 6$ and $\sup\{\tau : \bar{\bar{t}} < t(\mathbf{Z}, \mathbf{R} - \tau\mathbf{Z})\} = 5$, so the Hodges–Lehmann estimate is $\hat{\tau} = (6 + 5)/2 = 5.5$.

For particular test statistics, there are other ways of computing $\hat{\tau}$. This is true, for instance, for a single stratum using the rank sum test. In this case, it may be shown that $\hat{\tau}$ is the median of the $m(N - m)$ pairwise differences formed by taking each of the $m$ treated responses and subtracting each of the $N - m$ control responses.

The Hodges–Lehmann estimate $\hat{\tau}$ inherits properties from the test statistic $t(\cdot, \cdot)$. Consistency is one such property. Recall that a test is *consistent* if the probability of rejecting each false hypothesis tends to one as the sample size increases. Recall that an estimate is consistent if the probability that it is close to the true value tends to one as the sample size increases. As one would expect, these ideas are interconnected. A test that rejects incorrect values of $\tau$ leads to an estimate that moves away from these incorrect values. In other words, under mild conditions, consistent tests lead to consistent Hodges–Lehmann estimates; see Maritz (1981, §1.4) for some details.

## 2.8 *More Complex Outcomes

### 2.8.1 *Partially Ordered Outcomes

So far, the outcome $R_{si}$ has been a number, possibly a continuous measurement, possibly a binary event, possibly a discrete score, but always a single number. However, for more complex responses, much of the earlier discussion continues to apply with little or no change. The purpose of §2.8 is to discuss issues that arise with certain complex outcomes, including multivariate responses and censored observations.

When the outcome $R_{si}$ is a single number, it is clear what it means to speak of a high or low response, and it is clear what it means to ask whether responses are typically higher among treated units than among controls. For more complex responses, it may happen that some responses are higher than some others; and yet not every pair of possible responses can be ordered. For example, unit 1 may have a more favorable outcome than units 2 and 3, but units 2 and 3 may have different outcomes neither of which can be described as entirely more favorable than the other. For instance, patient 1 may live longer and have a better quality of life than patients 2 and 3, but patient 2 may outlive patient 3 though patient 3 had a better quality of life than patient 2. In this case, outcomes may be partially ordered rather than totally ordered, an idea that is formalized in a moment. Common examples are given in §2.8.2 and 2.8.3.

A *partially ordered set* or *poset* is a set $A$ together with a relation $\precsim$ on $A$ such that three conditions hold:

(i) $a \precsim a$ for all $a \in A$;

(ii) $a \precsim b$ and $b \precsim a$ implies $a = b$ for all $a, b \in A$; and

(iii) if $a \precsim b$ and $b \precsim c$ then $a \precsim c$ for all $a, b, c \in A$.

There is *strict inequality* between $a$ and $b$ if $a \precsim b$ and $a \neq b$. A poset $A$ is *totally ordered* if $a \precsim b$ or $b \precsim a$ for every $a, b \in A$. The real numbers with conventional inequality $\leq$ are totally ordered. If $A$ is partially ordered but not totally ordered, then for some $a, b \in A$, $a \neq b$, neither $a$ nor $b$ is higher than the other; that is, neither $a \precsim b$ nor $b \precsim a$. Sections 2.8.2 and 2.8.3 discuss two common examples of partially ordered outcomes, namely, censored and multivariate outcomes. Following this, in §2.8.4, general methods for partially ordered outcomes are discussed.

### 2.8.2 *Censored Outcomes

In some experiments, an outcome records the time to some event. In a clinical trial, the outcome may be the time between a patient's entry into the trial and the patient's death. In a psychological experiment, the outcome may be the time lapse between administration of a stimulus by the experimenter and the production of a response by an experimental subject. In a study of remedial education, the outcome may be the time until a certain level of proficiency in reading is reached.

Times may be censored in the sense that, when data analysis begins, the event may not yet have occurred. The patient may be alive at the close of the study. The stimulus may never elicit a response. The student may not develop proficiency in reading during the period under study.

If the event occurs for a unit after, say, 3 months, the unit's response is written 3. If the unit entered the study 3 months ago, if the event has not yet occurred, and if the analysis is done today, then the unit's response is written 3+ signifying that the event has not occurred in the initial 3 months.

Censored times are partially ordered. To see this, consider a simple illustration. In a clinical trial, patient 1 died at 3 months, patient 2 died at 12 months, and patient 3 entered the study 6 months ago and is alive today yielding a survival of 6+ months. Then patient 1 had a shorter survival than patients 2 and 3, but it is not possible to say whether patient 2 survived longer than patient 3 because we do not know whether patient 3 will survive for a full year.

The set $A$ of censored survival times contains the nonnegative real numbers together with the nonnegative real numbers with a plus appended. Define the partial order $\lesssim$ on $A$ as follows: if $a$ and $b$ are nonnegative real numbers, then:

(i) $a \lesssim b$ if and only if $a \leq b$;

(ii) $a \lesssim b+$ if and only if $a \leq b$; and

(iii) $a \lesssim a$ and $a+ \lesssim a+$.

Here, (i) indicates that "a" and "b" are both deaths and "a" died first. In (ii), "a" died before "b" was censored, so "b" certainly outlived "a." Of course, (iii) is just the case of equality—every censored time is equal to itself, and so is less than or equal to itself. It is easy to check that this is indeed a partial order, and that strict inequality indicates certainty about who died first.

### *2.8.3 *Multivariate Outcomes and Other Partially Ordered Outcomes*

Quite often, a single number is not enough to describe the outcome for a unit. In an educational intervention, there may be test scores in several areas, such as reading and mathematics. In a clinical trial, the outcome may involve both survival and quality of life. A multivariate response is a $p$-tuple of outcomes describing an individual. If the $p$ components are numbers, then the multivariate response inherits a partial order as follows: $(a_1, \ldots, a_p) \lesssim (b_1, \ldots, b_p)$ if and only if $a_1 \leq b_1, a_2 \leq b_2, \ldots$, and $a_p \leq b_p$. It is easy to check that this defines a partial order. As an example, if the outcome is the 2-tuple consisting of a reading score and a mathematics score, then one student has a higher multivariate response than another only if the first student did at least as well as the second student on both tests.

In fact, the components of the $p$-tuple need not be numbers—rather they may be any partially ordered outcomes. In the same way, the $p$-tuple inherits a partial order from the partial orders of individual outcomes. For instance, the outcome might be a 2-tuple consisting of a censored survival time and a number measuring quality of life. The censored survival times are partially but not totally ordered. In this case, a patient who died early with a poor quality of life would have a lower outcome than a patient who was censored late with a good quality of life.

Multivariate responses may be given other partial orders appropriate to particular contexts. Here is one that gives greatest emphasis to the first coordinate and about equal emphasis to the other two: $(a_1, a_2, a_3) \lesssim (b_1, b_2, b_3)$ if $a_1 \leq b_1$ or if $\{a_1 = b_1$ and $a_2 \leq b_2$ and $a_3 \leq b_3\}$. In an educational setting, this might say that a student who graduates had a better outcome than one who did not regardless of test scores, but among those who graduate, one student is better than another only if both reading and math scores are as good or better.

### 2.8.4 *A Test Statistic for Partially Ordered Outcomes

The task is to test the null hypothesis of no treatment effect against the alternative that treated units tend to have higher responses than controls in the sense of a partial order $\lesssim$ on the outcomes. For this purpose, define indicators $L_{sij}$ for $s = 1, \ldots, S$, $i = 1, \ldots, n_s$, $j = 1, \ldots, n_s$, as follows:

$$L_{sij} = \begin{cases} 1 & \text{if } R_{sj} \lesssim R_{si} \text{ with } R_{si} \neq R_{sj}, \\ -1 & \text{if } R_{si} \lesssim R_{sj} \text{ with } R_{si} \neq R_{sj}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

In words, $L_{sij}$ compares the $i$th and $j$th units in stratum $s$, and $L_{sij}$ is 1 if the $i$th is strictly greater than the $j$th, is $-1$ if the $i$th is strictly smaller than the $j$th, and is zero in all other cases. The statistic is

$$t(\mathbf{Z}, \mathbf{R}) = \sum_{s=1}^{S} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} Z_{si}(1 - Z_{sj}) L_{sij}. \tag{2.8}$$

Consider the statistic in detail. The term $Z_{si}(1 - Z_{sj})L_{sij}$ equals 1 if, in stratum $s$, the $i$th unit received the treatment, the $j$th unit received the control, and these two units had unequal responses with the treated unit having a higher response, $R_{sj} \lesssim R_{si}$. Similarly, $Z_{si}(1 - Z_{sj})L_{sij}$ equals $-1$ if, in stratum $s$, the $i$th unit is treated, the $j$th is a control, and the control had the higher response, $R_{si} \lesssim R_{sj}$. In all other cases, $Z_{si}(1 - Z_{sj})L_{sij}$ equals zero. So the test statistic is the number of comparisons of treated and control units in the same stratum in which the treated unit had the higher response minus the number in which the control unit had the higher response.

This statistic generalizes several familiar statistics. If the outcome is a single number and the partial order $\lesssim$ is ordinary inequality $\leq$, then (2.8) is equivalent to the Mann–Whitney (1947) statistic and the Wilcoxon (1945) rank sum statistic. If the outcome is censored and $\lesssim$ is the partial order in §2.8.2, then the statistic is Gehan's (1965) statistic.

A device due to Mantel (1967) shows that (2.8) is a sum statistic. The steps are as follows. First note that, for any subset $B$ of $\{1, 2, \ldots, n_s\}$,

$$\sum_{i \in B} \sum_{j \in B} L_{sij} = 0 \tag{2.9}$$

since $L_{sij}$ and $L_{sji}$ both appear in the sum, with $L_{sij} = -L_{sji}$, and they cancel. Using this fact with $B = \{i : 1 \leq i \leq n_s \text{ with } Z_{si} = 1\}$ yields

$$0 = \sum_{i \in B} \sum_{j \in B} L_{sij} = \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} Z_{si} Z_{sj} L_{sij},$$

which permits the test statistic (2.8) to be rewritten as the sum statistic

$$t(\mathbf{Z}, \mathbf{R}) = \sum_{s=1}^{S} \sum_{i=1}^{n_s} Z_{si} \sum_{j=1}^{n_s} L_{sij} = \sum_{s=1}^{S} \sum_{i=1}^{n_s} Z_{si} q_{si} \quad \text{with} \quad q_{si} = \sum_{j=1}^{n_s} L_{sij}.$$

As a result, the expectation and variance of the test statistic under the null hypothesis are given by Proposition 2. In fact, in that Proposition, $\bar{q}_s = 0$ for each $s$ using (2.9).

The score $q_{si}$ has an interpretation. It is the number of units in stratum $s$ with outcomes less than unit $i$ minus the number with outcomes greater than $i$. The score $q_{si}$ is large if unit $i$ has a response larger than that of most units in stratum $s$. For instance, in Gehan's statistic for censored outcomes, the score $q_{si}$ is the number of patients in stratum $s$ who definitely died before patient $i$ minus the number who definitely died after patient $i$.

### *2.8.5 *Effect Increasing Statistics, Positive Effects, Larger Effects*

In §2.4 and 2.5, three terms were discussed, namely, effect increasing statistics, positive effects, and larger effects. These terms apply to partially ordered outcomes with virtually no change, as shown in a moment. In each case, the definitions in §2.4 and 2.5 are the special case of the definitions in this section with the partial order $\lesssim$ given by ordinary inequality $\leq$ of real numbers.

Let $\mathbf{r}$ and $\mathbf{r}^*$ be two possible values of the $N$-tuple of partially ordered outcomes. If $r_{si} \lesssim r^*_{si}$ for every treated unit and $r^*_{si} \lesssim r_{si}$ for every control unit, then the treated and control groups appear farther apart for outcome $\mathbf{r}^*$ than for outcome $\mathbf{r}$. A test statistic $t(\cdot,\cdot)$ is *effect increasing* if $t(\mathbf{z},\mathbf{r}) \leq t(\mathbf{z},\mathbf{r}^*)$ whenever $\mathbf{r}$ and $\mathbf{r}^*$ are two possible values of the response such that $r_{si} \lesssim r^*_{si}$ if $z_{si} = 1$ and $r^*_{si} \lesssim r_{si}$ if $z_{si} = 0$ for all $s$, $i$. In words, the statistic is larger when the outcomes in treated and control groups are farther apart. The statistic in §2.8.4 is effect increasing; see Problem 6.

If there is no interference between units, then $(\mathbf{r}_T, \mathbf{r}_C)$ is a *positive effect* if $\mathbf{r}_T \neq \mathbf{r}_C$ and $r_{Csi} \lesssim r_{Tsi}$ for every $s$, $i$. In the case of censored survival times, this would mean that each patient would definitely survive at least as long under the treatment as under the control, or else would continue to be censored at the same time due to the end of the study. An effect $(\mathbf{r}^*_T, \mathbf{r}^*_C)$ is a *larger effect* than $(\mathbf{r}_T, \mathbf{r}_C)$ if $r_{Tsi} \lesssim r^*_{Tsi}$ and $r^*_{Csi} \lesssim r_{Csi}$, for all $s$, $i$, that is, if the treated responses are higher and the control responses are lower.

The following proposition is the extension of Proposition 4 to partially ordered responses. Again, the proof is given in the appendix, §2.9.

**Proposition 5** *In a randomized experiment, a test statistic that is effect increasing yields an unbiased test of no effect against the alternative of a positive effect, and if $(\mathbf{r}^*_T, \mathbf{r}^*_C)$ is a larger effect than $(\mathbf{r}_T, \mathbf{r}_C)$ then $t(\mathbf{Z},\mathbf{R}^*) \geq t(\mathbf{Z},\mathbf{R})$.*

## 2.9 *Appendix: Effect Increasing Tests Under Alternatives

This appendix proves Propositions 4 and 5 which describe the behavior of effect increasing test statistics under the alternative hypotheses of positive effects or larger effects. It may be of interest to contrast these propositions with a result in Lehmann (1959, §5.8, Lemma 2) which is similar in spirit though quite different in detail. It suffices to prove Proposition 5 since Proposition 4 is the special case of the former in which the partial order is ordinary inequality. The proof depends on the following lemma.

**Lemma 6** *Let $t(\cdot,\cdot)$ be effect increasing. If $(\mathbf{r}_T, \mathbf{r}_C)$ is a positive effect, then $t(\mathbf{z},\mathbf{r}_\mathbf{z}) \geq t(\mathbf{z},\mathbf{r}_\mathbf{a})$ for all $\mathbf{z}$, $\mathbf{a} \in \Omega$. If $(\mathbf{r}^*_T, \mathbf{r}^*_C)$ is a larger effect than $(\mathbf{r}_T, \mathbf{r}_C)$, then $t(\mathbf{z},\mathbf{r}^*_\mathbf{z}) \geq t(\mathbf{z},\mathbf{r}_\mathbf{z})$ for all $\mathbf{z} \in \Omega$.*

**Proof of Lemma.** Let $(\mathbf{r}_T, \mathbf{r}_C)$ be a positive effect, let $\mathbf{z}$, $\mathbf{a} \in \Omega$, and consider $\mathbf{r}_\mathbf{z}$ and $\mathbf{r}_\mathbf{a}$. If $z_{si} = 1$, then $r_{si\mathbf{z}} = \mathbf{r}_{Tsi}$ while $\mathbf{r}_{si\mathbf{a}}$ may equal either

$\mathbf{r}_{Tsi}$ or $\mathbf{r}_{Csi}$ depending on $a_{si}$, but in either case $r_{si\mathbf{a}} \lesssim r_{si\mathbf{z}}$ since $(\mathbf{r}_T, \mathbf{r}_C)$ is a positive effect. Similarly, if $z_{si} = 0$, then $r_{si\mathbf{z}} = r_{Csi} \lesssim r_{si\mathbf{a}}$. Since $t(\cdot,\cdot)$ is effect increasing, this implies $t(\mathbf{z}, \mathbf{r_z}) \geq t(\mathbf{z}, \mathbf{r_a})$, proving the first part of the lemma.

Now let $\mathbf{z} \in \Omega$, let $(\mathbf{r}_T^*, \mathbf{r}_C^*)$ be a larger effect than $(\mathbf{r}_T, \mathbf{r}_C)$, and consider $\mathbf{r}_\mathbf{z}^*$ and $\mathbf{r_z}$. If $z_{si} = 1$, then $r_{si\mathbf{z}} = r_{Tsi} \lesssim r_{Tsi}^* = r_{si\mathbf{z}}^*$. If $z_{si} = 0$, then $r_{si\mathbf{z}}^* = r_{Csi}^* \lesssim r_{Csi} = r_{si\mathbf{z}}$. Hence $t(\mathbf{z}, \mathbf{r}_\mathbf{z}^*) \geq t(\mathbf{z}, \mathbf{r_z})$ since $t(\cdot,\cdot)$ is effect increasing, completing the proof. ■

**Proof of Proposition 5.** The lemma directly shows that if $(\mathbf{r}_T^*, \mathbf{r}_C^*)$ is a larger effect than $(\mathbf{r}_T, \mathbf{r}_C)$, then $t(\mathbf{Z}, \mathbf{R}^*) \geq t(\mathbf{Z}, \mathbf{R})$. To prove unbiasedness, let $\mathbf{Z}$ be randomly selected from $\Omega$ where $\text{prob}(\mathbf{Z} = \mathbf{z})$ is known but need not be uniform. If the random treatment assignment turns out to be $\mathbf{Z} = \mathbf{a}$, then the observed outcome is $\mathbf{R} = \mathbf{r_a}$. If the null hypothesis were true, if the treatment had no effect, the observed response would be the same $\mathbf{r_a}$ no matter how treatments were assigned, that is, the observed response would be $\mathbf{R} = \mathbf{r_a}$ no matter what value $\mathbf{Z}$ assumed. If the null hypothesis were false and the treatment had a positive effect, the observed response would vary depending upon the treatment assignment, $\mathbf{R} = \mathbf{r_z}$ if $\mathbf{Z} = \mathbf{z}$. For any fixed number $T$

$$\begin{aligned}
&\text{prob}\{t(\mathbf{Z}, \mathbf{R}) \geq T\} \\
&\quad = \sum_{\mathbf{z} \in \Omega} [t(\mathbf{z}, \mathbf{r_z}) \geq T] \;\; \text{prob}(\mathbf{Z} = \mathbf{z}) \\
&\quad \geq \sum_{\mathbf{z} \in \Omega} [t(\mathbf{z}, \mathbf{r_a}) \geq T] \;\; \text{prob}(\mathbf{Z} = \mathbf{z}) \qquad \text{for } \mathbf{a} \in \Omega \text{ by the lemma.}
\end{aligned}$$

In other words, the chance that the test statistic $t(\mathbf{Z}, \mathbf{R})$ exceeds any number $T$ is at least as great under the alternative hypothesis of a positive effect as under the null hypothesis of no effect, proving unbiasedness. ■

## 2.10 *Appendix: The Set of Treatment Assignments

### *2.10.1 *Outline and Motivation: The Special Structure of Ω*

The set $\Omega$ of treatment assignments plays an important role both in randomized experiments and in the discussion of observational studies in later chapters. This set $\Omega$ possess a special structure, first noted by Savage (1964). Using this structure, a single theorem may refer to large classes of test statistics and to all of the simple designs, including matched pairs, matching with multiple controls, two-group comparisons, and stratified comparisons. The purpose of this section is to describe the special structure of $\Omega$. Appendices in later chapters refer back to this appendix.

Savage (1964) observed that the set $\Omega$ is a finite distributive lattice. This is useful because there are tidy theorems about probability distributions on a finite distributive lattice, including the FKG inequality and Holley's inequality. This section:

(i) offers a little motivation;

(ii) reviews the definition of a distributive lattice;

(iii) shows that $\Omega$ is indeed such a lattice; and

(iv) discusses the relevant probability inequalities.

The material in this appendix may be read without previous experience with lattices.

For motivation, consider a simple case. There is a single stratum, $S = 1$, so the $s$ subscript is dropped in this example, and there are $n = 4$ units of which $m = 2$ receive the treatment. Then $\Omega$ contains $\binom{4}{2} = 6$ possible treatment assignments. Assume for this motivating example that the null hypothesis of no treatment effect holds, and renumber the four subjects so their observed responses are in decreasing order, $r_1 \geq r_2 \geq r_3 \geq r_4$. Since no quantity we calculate ever depends on the numbering of subjects, this renumbering changes nothing, but it is notationally convenient. The six possible treatment assignments appear in (2.10).

$$\begin{tikzcd}
& 1100 & \\
& 1010 \arrow[u] & \\
0110 \arrow[ur] & & 1001 \arrow[ul] \\
& 0101 \arrow[ul] \arrow[ur] & \\
& 0011 \arrow[u] &
\end{tikzcd} \tag{2.10}$$

The treatment assignment $\mathbf{z} = (1, 1, 0, 0)$ at the top in (2.10) is the one that would suggest the largest positive treatment effect, since this assignment places the two largest responses, $r_1$ and $r_2$, in the treated group. The assignment below this, namely, $\mathbf{z} = (1, 0, 1, 0)$ would suggest a smaller treatment effect than (1, 1, 0, 0), since $r_3$ has replaced $r_2$, but it would suggest a larger treatment effect than any other assignment. The assignments (0, 1, 1, 0) and (1, 0, 0, 1) are not directly comparable to each other, since the latter places the largest and smallest responses in the treated group while the former places the two middle responses in the treated group; however, both are lower than (1, 0, 1, 0) and both are higher than (0, 1, 0, 1).

Consider the behavior of a test statistic $t(\mathbf{z}, \mathbf{r})$ as we move through (2.10). Suppose, for instance, there are no ties among the responses, $r_1 > r_2 >$

$r_3 > r_4$, and $t(\mathbf{z}, \mathbf{r})$ is the rank sum statistic. Then $t(\mathbf{z}, \mathbf{r}) = 7$ for $\mathbf{z} = 1100$, $t(\mathbf{z}, \mathbf{r}) = 6$ for 1010, $t(\mathbf{z}, \mathbf{r}) = 5$ for both 1001 and 0110, $t(\mathbf{z}, \mathbf{r}) = 4$ for 0101, and $t(\mathbf{z}, \mathbf{r}) = 3$ for 0011, so $t(\mathbf{z}, \mathbf{r})$ increases steadily along upward paths in (2.10). If, instead, $t(\mathbf{z}, \mathbf{r})$ were the difference between the mean response in treated and control groups, it would again be increasing along upward paths.

Suppose, instead, that $r_2$ and $r_3$ were tied, so $r_1 > r_2 = r_3 > r_4$. In this case, the rank sum statistic would give average rank 2.5 to both $r_2$ and $r_3$, so moving from 1100 to 1010 would not change $t(\mathbf{z}, \mathbf{r})$. Notice, however, that even with ties, $t(\mathbf{z}, \mathbf{r})$ is monotone increasing (i.e., nondecreasing) along upward paths.

Actually, the order in (2.10) applies to many statistics whether ties are present or not. If $t(\mathbf{z}, \mathbf{r})$ is any arrangement-increasing statistic, then $t(\mathbf{z}, \mathbf{r})$ is monotone-increasing on upward paths in (2.10). Most reasonable statistics will assign a higher value to 1100 than to 1010, but reasonable statistics can differ in how they order assignments that are not comparable like 1001 and 0110.

Take a look at a second example, the case of $S = 2$ matched pairs, so $n_s = 2$ and $m_s = 1$ for $s = 1, 2$. Then $\Omega$ contains $2^2 = 4$ treatment assignments $\mathbf{z} = (z_{11}, z_{12}, z_{21}, z_{22})$. Again, assume the null hypothesis of no treatment effect and renumber the units in each pair so that in the first pair $r_{11} \geq r_{12}$, and in the second pair $r_{21} \geq r_{22}$. The set $\Omega$ appears in (2.11).

$$\begin{tikzcd}
 & 1010 & \\
1001 \arrow[ur] & & 0110 \arrow[ul] \\
 & 0101 \arrow[ul] \arrow[ur] &
\end{tikzcd} \tag{2.11}$$

The assignment $\mathbf{z}$ in (2.11) suggesting the largest positive treatment effect is $\mathbf{z} = (1, 0, 1, 0)$ since in both pairs the treated unit had a higher response than the control. For $\mathbf{z} = 1001$ and $\mathbf{z} = 0110$, the treated unit had the higher response in one pair and the lower response in the other. In the assignment $\mathbf{z} = 0101$ the treated unit had a lower response than the control in both pairs.

Once again, common statistics are monotone-increasing along upward paths in (2.11). For instance, this is true of the signed rank statistic, which equals zero at the bottom of (2.11), equals one or two in the middle, and equals three at the top. Indeed, all arrangement-increasing functions are monotone-increasing along upward paths in (2.11).

What does all this suggest? There are certain treatment assignments $\mathbf{z} \in \Omega$ that are higher than others, and this is true without reference to the nature of the response $\mathbf{r}$ or the specific test statistic $t(\mathbf{z}, \mathbf{r})$. The responses might be continuous or they might be discrete scores or they might be binary. The test statistic might be the signed rank statistic or the McNemar statistic. In all these cases, $\mathbf{z} = 1010$ is higher than $\mathbf{z} = 1001$ in (2.11). Certain statements about treatment assignments $\mathbf{z} \in \Omega$ should be true generally, without reference to the specific nature of the outcome or the test statistic.

### 2.10.2 *A Brief Review of Lattices

Briefly, a lattice is a partially ordered set in which each pair of elements has a greatest lower bound and a least upper bound. This terminology is discussed formally in a moment, but first consider what this means in (2.10). A point $\mathbf{z}$ in (2.10) is below another $\mathbf{z}^*$ if there is a path up from $\mathbf{z}$ to $\mathbf{z}^*$; for instance, 0110 is below 1100. The points 1001 and 0110 are not comparable—there is not a path up from one to the other—so $\Omega$ is partially but not totally ordered. The least upper bound of 0110 and 1001 is 1010, for it is the smallest element above both of them. The least upper bound of 1010 and 1100 is 1100. A nice introduction to lattices is given by MacLane and Birkoff (1988).

A set $\Omega$ is *partially ordered* by a relation $\lesssim$ if for all $\mathbf{z}, \mathbf{z}^*, \mathbf{z}^{**} \in \Omega$:

(i) $\mathbf{z} \lesssim \mathbf{z}$;

(ii) $\mathbf{z} \lesssim \mathbf{z}^*$ and $\mathbf{z}^* \lesssim \mathbf{z}$ implies $\mathbf{z} = \mathbf{z}^*$; and

(iii) $\mathbf{z} \lesssim \mathbf{z}^*$ and $\mathbf{z}^* \lesssim \mathbf{z}^{**}$ implies $\mathbf{z} \lesssim \mathbf{z}^{**}$.

An *upper bound* for $\mathbf{z}, \mathbf{z}^* \in \Omega$ is an element $\mathbf{z}^{**}$ such that $\mathbf{z} \lesssim \mathbf{z}^{**}$ and $\mathbf{z} \lesssim \mathbf{z}^{**}$. A *least upper bound* $\mathbf{z}^{**}$ for $\mathbf{z}, \mathbf{z}^*$ is an upper bound that is below all other upper bounds for $\mathbf{z}, \mathbf{z}^*$; that is, if $\mathbf{z}^{***}$ is any upper bound for $\mathbf{z}, \mathbf{z}^*$, then $\mathbf{z}^{**} \lesssim \mathbf{z}^{***}$. If a least upper bound for $\mathbf{z}, \mathbf{z}^*$ exists, then it is unique by (ii). Lower bound and *greatest lower bound* are defined similarly. A *lattice* is a partially ordered set $\Omega$ in which every pair $\mathbf{z}, \mathbf{z}^*$ of elements has a least upper bound, written $\mathbf{z} \vee \mathbf{z}^*$, and a greatest lower bound, written $\mathbf{z} \wedge \mathbf{z}^*$. A lattice $\Omega$ is *finite* if the set $\Omega$ contains only finitely many elements. In (2.10), both 1010 and 1100 are upper bounds for the pair 1001 and 0110, but the least upper bound is $1001 \vee 0110 = 1010$.

The partial order $\lesssim$ and the operations $\vee$ and $\wedge$ are tied together by the following relationship: $\mathbf{z} \lesssim \mathbf{z}^*$ if and only if $\mathbf{z} \vee \mathbf{z}^* = \mathbf{z}^*$ and $\mathbf{z} \wedge \mathbf{z}^* = \mathbf{z}$. In fact, using this relationship, a lattice may be defined beginning with the operations $\vee$ and $\wedge$ rather than beginning with the partial order $\lesssim$, that is, defining the partial order in terms of the operations. The following theorem is well known; see MacLane and Birkoff (1988, §XIV, 2) for proof.

**Theorem 7** *A set $\Omega$ with operations $\vee$ and $\wedge$ is a lattice if and only if for all $\mathbf{z}, \mathbf{z}^*, \mathbf{z}^{**} \in \Omega$:*

*L1.* $\mathbf{z} \vee \mathbf{z} = \mathbf{z}$ *and* $\mathbf{z} \wedge \mathbf{z} = \mathbf{z}$;

*L2.* $\mathbf{z} \vee \mathbf{z}^* = \mathbf{z}^* \vee \mathbf{z}$ *and* $\mathbf{z} \wedge \mathbf{z}^* = \mathbf{z}^* \wedge \mathbf{z}$;

*L3.* $\mathbf{z} \vee (\mathbf{z}^* \vee \mathbf{z}^{**}) = (\mathbf{z} \vee \mathbf{z}^*) \vee \mathbf{z}^{**}$ *and* $\mathbf{z} \wedge (\mathbf{z}^* \wedge \mathbf{z}^{**}) = (\mathbf{z} \wedge \mathbf{z}^*) \wedge \mathbf{z}^{**}$; *and*

*L4.* $\mathbf{z} \wedge (\mathbf{z} \vee \mathbf{z}^*) = \mathbf{z} \vee (\mathbf{z} \wedge \mathbf{z}^*) = \mathbf{z}$.

Here, L2 and L3 are the commutative and associate laws, L1 is called idempotence, and L4 is called absorption. A lattice is *distributive* if the distributive law also holds,

$$\mathbf{z} \vee (\mathbf{z}^* \wedge \mathbf{z}^{**}) = (\mathbf{z} \vee \mathbf{z}^*) \wedge (\mathbf{z} \vee \mathbf{z}^{**}) \qquad \text{for all} \quad \mathbf{z}, \mathbf{z}^*, \mathbf{z}^{**} \in \Omega.$$

### 2.10.3 *The Set of Treatment Assignments Is a Distributive Lattice

This section gives Savage's (1964) demonstration that $\Omega$ is a distributive lattice. With each $N$-dimensional $\mathbf{z} \in \Omega$, associate a vector $\mathbf{c}$ of dimension $\sum m_s$, as follows. The vector $\mathbf{c}$ is made up of $S$ pieces, where piece $s$ has $m_s$ coordinates. It is suggestive and almost accurate to say that $\mathbf{c}$ contains the ranks of the responses of treated units, each stratum being ranked separately, the ranks being arranged in decreasing order in each stratum. This would be exactly true if there were no ties, but it is not exactly true in the case of ties. Here is the exact definition, with or without ties. If $z_{s1} = 0$, $z_{s2} = 0, \ldots, z_{s,i-1} = 0$, $z_{si} = 1$, then $c_{s1} = n_s - i + 1$. Continuing, if $z_{s,i+1} = 0, \ldots, z_{s,j-1} = 0$, $z_{sj} = 1$, then $c_{s2} = n_s - j + 1$, and so on. In terms of the $\mathbf{c}$, (2.10) becomes (2.12), and (2.11) becomes (2.13). For instance, in (2.10), $\mathbf{z} = 1100$ becomes $\mathbf{c} = 43$, since the first 1 in $\mathbf{z}$ appears in position $i = 1$, so $n - i + 1 = 4 - 1 + 1 = 4$ and the second 1 in $\mathbf{z}$ appears in position $j = 2$, so $n - j + 1 = 4 - 2 + 1 = 3$.

$$\begin{tikzcd}
 & 43 & \\
 & 42 \arrow[u] & \\
41 \arrow[ur] & & 32 \arrow[ul] \\
 & 31 \arrow[ul] \arrow[ur] & \\
 & 21 \arrow[u] &
\end{tikzcd}$$

(2.12)

If there are ties among the responses in a stratum, then $\mathbf{c}$ is no longer a collection of ranks, because $\mathbf{c}$ distinguishes units with the same tied response. In the end, this is not a problem. The lattice order makes a few distinctions among treatment assignments that statistical procedures will ignore.

$$\begin{tikzcd}
 & 22 & \\
21 \arrow[ur] & & 12 \arrow[ul] \\
 & 11 \arrow[ul] \arrow[ur] &
\end{tikzcd} \tag{2.13}$$

It is readily checked that each $\mathbf{z}$ has one and only one corresponding $\mathbf{c}$. Given $\mathbf{z}$, $\mathbf{z}^* \in \Omega$, with corresponding $\mathbf{c}$ and $\mathbf{c}^*$, the operations $\vee$ and $\wedge$ are defined as follows. Define $\mathbf{c} \vee \mathbf{c}^*$ and $\mathbf{c} \wedge \mathbf{c}^*$ as the vectors containing, respectively, $\max(c_{si}, c^*_{si})$ and $\min(c_{si}, c^*_{si})$. Define $\mathbf{z} \vee \mathbf{z}^*$ and $\mathbf{z} \wedge \mathbf{z}^*$ as the elements of $\Omega$ corresponding to $\mathbf{c} \vee \mathbf{c}^*$ and $\mathbf{c} \wedge \mathbf{c}^*$. It is readily checked that this definition makes sense, that is, that $\mathbf{c} \vee \mathbf{c}^*$ and $\mathbf{c} \wedge \mathbf{c}^*$ always correspond to elements of $\Omega$. For instance, in (2.10), $\mathbf{z} = 0110$ and $\mathbf{z}^* = 1001$ correspond to $\mathbf{c} = 32$ and $\mathbf{c}^* = 41$, so $\mathbf{c} \vee \mathbf{c}^* = 42$ and $\mathbf{c} \wedge \mathbf{c}^* = 31$, so $\mathbf{z} \vee \mathbf{z}^* = 1010$ and $\mathbf{z} \wedge \mathbf{z}^* = 0101$, as is consistent with (2.10). Notice carefully that the coordinate $(s, i)$ of $\mathbf{z} \vee \mathbf{z}^*$ is not generally equal to $\max(z_{si}, z^*_{si})$.

To show that $\Omega$ is a lattice with these operations, one needs to check L1 to L4 in Theorem 7, but L1 to L3 hold trivially for $\max(c_{si}, c^*_{si})$ and $\min(c_{si}, c^*_{si})$. To show $\mathbf{z} \wedge (\mathbf{z} \vee \mathbf{z}^*) = \mathbf{z}$ in L4, it suffices to show $\mathbf{c} \wedge (\mathbf{c} \vee \mathbf{c}^*) = \mathbf{c}$. If $c_{si} \geq c^*_{si}$, then $\min\{c_{si}, \max(c_{si}, c^*_{si})\} = \min(c_{si}, c_{si}) = c_{si}$, while if $c_{si} < c^*_{si}$, then $\min\{c_{si}, \max(c_{si}, c^*_{si})\} = \min(c_{si}, c^*_{si}) = c_{si}$, so $\mathbf{c} \wedge (\mathbf{c} \vee \mathbf{c}^*) = \mathbf{c}$ as required. The second part of L4 is proved in the same way. So $\Omega$ is a lattice.

More than this, $\Omega$ is a distributive lattice. As proof, it suffices to show $\mathbf{c} \vee (\mathbf{c}^* \wedge \mathbf{c}^{**}) = (\mathbf{c} \vee \mathbf{c}^*) \wedge (\mathbf{c} \vee \mathbf{c}^{**})$, that is, to show

$$\max\{c_{si}, \min(c^*_{si}, c^{**}_{si})\} = \min\{\max(c_{si}, c^*_{si}), \max(c_{si}, c^{**}_{si})\}.$$

There are two cases. If $c_{si} \geq \min(c^*_{si}, c^{**}_{si})$, then $\max\{c_{si}, \min(c^*_{si}, c^{**}_{si})\} = c_{si}$, but also $c_{si}$ is less than or equal to both $\max(c_{si}, c^*_{si})$ and $\max(c_{si}, c^{**}_{si})$ yet it equals one of them, so

$$\min\{\max(c_{si}, c^*_{si}), \max(c_{si}, c^{**}_{si})\} = c_{si}.$$

On the other hand, if $c_{si} < \min(c^*_{si}, c^{**}_{si})$, then

$$\max\{c_{si}, \min(c^*_{si}, c^{**}_{si})\} = \min(c^*_{si}, c^{**}_{si}),$$

but $\max(c_{si}, c^*_{si}) = c^*_{si}$, and $\max(c_{si}, c^{**}_{si}) = c^{**}_{si}$, so

$$\min\{\max(c_{si}, c^*_{si}), \max(c_{si}, c^{**}_{si})\} = \min(c^*_{si}, c^{**}_{si}),$$

as required to complete the proof.

### 2.10.4 *Inequalities for Probability Distributions on a Lattice

This section discusses two inequalities for probability distributions on a finite distributive lattice, namely, the FKG inequality and Holley's inequality. These inequalities are the principal tool that makes use of the lattice properties of $\Omega$. The original proofs of these inequalities are somewhat involved, but Ahlswede and Daykin (1978) developed a simpler proof involving nothing more than elementary probability. Their proof is nicely presented in several recent texts (Anderson 1987, §6, Bollobas, 1986, §19), to which the reader may refer.

A real-valued function on $\Omega$, $f : \Omega \rightarrow \mathbb{R}$ is isotonic if $\mathbf{z} \precsim \mathbf{z}^*$ implies $f(\mathbf{z}) \leq f(\mathbf{z}^*)$. Throughout this appendix, $\mathbf{r}$ has been sorted into order within each stratum, $r_{si} \geq r_{s,i+1}$ for each $s$, $i$. With this order, the arrangement-increasing statistics $t(\mathbf{z}, \mathbf{r})$ are some of the isotonic functions on $\Omega$. Actually, the arrangement-increasing statistics are the interesting isotonic functions, for they are the isotonic functions that are unchanged by interchanging tied responses in the same stratum. If there are ties, that is, if $r_{si} = r_{s,i+1}$ for some $s$ and $i$, then there are isotonic functions that are not arrangement-increasing, specifically functions that increase when $z_{si} = 0, z_{s,i+1} = 1$ is replaced by $z_{si} = 1, z_{s,i+1} = 0$; however, these functions are not interesting as test statistics $t(\mathbf{z}, \mathbf{r})$ because they distinguish between people who gave identical responses. From a practical point of view, the important point is that a property of all isotonic functions on $\Omega$ is automatically a property of all arrangement-increasing functions, and all of the statistics in §2.4.3 are arrangement-increasing.

The first inequality is due to Fortuin, Kasteleyn, and Ginibre (1971).

**Theorem 8 (The FKG Inequality)** *Let $f(\cdot)$ and $g(\cdot)$ be isotonic functions on a finite distributive lattice $\Omega$. If a random element $\mathbf{Z}$ of $\Omega$ is selected by a probability distribution satisfying*

$$\text{prob}(\mathbf{Z} = \mathbf{z} \vee \mathbf{z}^*) \cdot \text{prob}(\mathbf{Z} = \mathbf{z} \wedge \mathbf{z}^*) \geq \text{prob}(\mathbf{Z} = \mathbf{z}) \cdot \text{prob}(\mathbf{Z} = \mathbf{z}^*)$$
$$\textit{for all} \quad \mathbf{z}, \mathbf{z}^* \in \Omega,$$

*then*

$$\text{cov}\{f(\mathbf{Z}), g(\mathbf{Z})\} \geq 0.$$

For example, randomization gives equal probabilities to all elements of $\Omega$, so the randomization distribution satisfies the condition for the FKG inequality. Hence, under the null hypothesis of no effect in a randomized experiment, any two arrangement-increasing statistics have a nonnegative correlation.

The next theorem is due to Holley (1974).

**Theorem 9 (Holley's Inequality)** *Let $f(\cdot)$ be an isotonic function on a finite distributive lattice $\Omega$. If $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are random elements of $\Omega$ selected* *by two probability distributions satisfying*

$$\text{prob}(\mathbf{Z} = \mathbf{z} \vee \mathbf{z}^*) \cdot \text{prob}(\tilde{\mathbf{Z}} = \mathbf{z} \wedge \mathbf{z}^*) \geq \text{prob}(\mathbf{Z} = \mathbf{z}) \cdot \text{prob}(\tilde{\mathbf{Z}} = \mathbf{z}^*)$$
$$\textit{for all} \quad \mathbf{z}, \mathbf{z}^* \in \Omega,$$

*then*

$$E\{f(\mathbf{Z})\} \geq E\{f(\tilde{\mathbf{Z}})\}.$$

In other words, the premise of Holley's inequality is a sufficient condition for $\mathbf{Z}$ to be stochastically larger than $\tilde{\mathbf{Z}}$, in the sense that for every arrangement-increasing function $f(\cdot)$, the random variable $f(\mathbf{Z})$ has higher expectation than $f(\tilde{\mathbf{Z}})$. Holley's inequality helps later in comparing a nonrandom assignment of treatments to a random assignment. A related result is given by Krieger and Rosenbaum (1994). Literature related to Holley's inequality is reviewed in Rosenbaum (1999).

### *2.10.5 *An Identity in $\Omega$*

There is a useful identity in the set $\Omega$ of treatment assignments. The identity links $\vee$ and $\wedge$ to the addition of vectors, and therefore it is useful in verifying the conditions of the FKG inequality and Holley's inequality. It is true for this lattice, but not true generally for all lattices.

**Lemma 10** *For all* $\mathbf{z}, \mathbf{z}^* \in \Omega$,

$$\mathbf{z} \vee \mathbf{z}^* + \mathbf{z} \wedge \mathbf{z}^* = \mathbf{z} + \mathbf{z}^*.$$

**Proof.** Fix a coordinate $(s, i)$, so the task is to show $z_{si} + z^*_{si} = z_{\wedge si} + z_{\vee si}$, where $z_{\wedge si}$ and $z_{\vee si}$ are the $(s, i)$ coordinates of $\mathbf{z} \wedge \mathbf{z}^*$ and $\mathbf{z} \vee \mathbf{z}^*$, respectively. Let $\mathbf{c}$ and $\mathbf{c}^*$ correspond with $\mathbf{z}$ and $\mathbf{z}^*$, respectively. There are three cases, depending upon the value of $z_{si} + z^*_{si}$.

1. If $z_{si} + z^*_{si} = 0$, then $c_{sj} \neq n_s - i + 1$ and $c^*_{sj} \neq n_s - i + 1$ for $j = 1, \ldots, m_s$, so $\max\left(c_{sj}, c^*_{sj}\right) \neq n_s - i + 1$ and $\min\left(c_{sj}, c^*_{sj}\right) \neq n_s - i + 1$ for $j = 1, \ldots, m_s$, so $z_{\wedge si} + z_{\vee si} = 0$, as required.

2. If $z_{si} + z^*_{si} = 2$, then there is a $j$ and a $k$ such that $c_{sj} = n_s - i + 1$ and $c^*_{sk} = n_s - i + 1$. If $j = k$, then $\max\left(c_{sj}, c^*_{sj}\right) = n_s - i + 1$ and $\min\left(c_{sj}, c^*_{sj}\right) = n_s - i + 1$, so $z_{\wedge si} = 1$ and $z_{\vee si} = 1$, so that $z_{\wedge si} + z_{\vee si} = 2$, as required. If $j < k$, then $n_s - i + 1 = c_{sj} > c_{sk}$ and $c^*_{sj} > c^*_{sk} = n_s - i + 1$, so $\min\left(c_{sj}, c^*_{sj}\right) = c_{sj} = n_s - i + 1$ and $\max\left(c_{sk}, c^*_{sk}\right) = c^*_{sk} = n_s - i + 1$, so $z_{\wedge si} = 1$ and $z_{\vee si} = 1$, so that $z_{\wedge si} + z_{\vee si} = 2$, as required. The case $j > k$ is similar.

3. If $z_{si} = 1$ and $z^*_{si} = 0$, so $z_{si} + z^*_{si} = 1$, then there is a $j$ such that $c_{sj} = n_s - i + 1$ but $c^*_{sk} \neq n_s - i + 1$ for $k = 1, \ldots, m_s$. In this case, either $n_s - i + 1 = \max\left(c_{sj}, c^*_{sj}\right)$ or $n_s - i + 1 = \min\left(c_{sj}, c^*_{sj}\right)$ but not both, and moreover, $n_s - i + 1 \neq \max\left(c_{sk}, c^*_{sk}\right)$ and $n_s - i + 1 \neq \min\left(c_{sk}, c^*_{sk}\right)$ for all

$k \neq j$, so $z_{\wedge si} + z_{\vee si} = 1$, as required. The case $z_{si} = 0$ and $z^*_{si} = 1$ is similar. ■

If there were no ties, so **c** and **c*** are ranks, then Lemma 10 has the following interpretation. Within each stratum, the operations $\vee$ and $\wedge$ take the ranks in **c** and **c*** and apportion them in forming $\mathbf{c} \vee \mathbf{c}^*$ and $\mathbf{c} \wedge \mathbf{c}^*$, but in this process they do not create or delete ranks that appear in **c** and **c***.

## 2.11 Bibliographic Notes

Fisher is usually credited with the invention of randomized experiments. See, in particular, his important and influential book, *The Design of Experiments*, first published in 1935. Randomization is discussed in many articles and textbooks. In particular, see Kempthorne (1952), Cox (1958a, §5) and Cox and Reid (2000) for discussions of randomization in experimental design, and see Lehmann (1975) and Maritz (1981) for discussions of its role in nonparametrics. Mantel's (1963) paper was significant not just for the method he proposed, but also for its strengthening of the link between nonparametric methods and contingency table methods. The model for a treatment effect in §2.5.2 in which each unit has two potential responses, one under treatment and the other under control, has a long history. In an article first published in Polish and recently translated into English, Neyman (1923) used it to study the behavior of statistical tests under random assignment of treatments. Related work was done by Welch (1937), Wilk (1955), Cox (1958b, §5), and Robinson (1973), among others. Rubin (1974, 1977) first used the model in observational studies. In particular, he discussed the conditions under which matching, stratification, and covariance adjustment all estimate the same treatment effect. See also Hamilton (1979) and Holland (1986). Arrangement-increasing functions have been studied under various names by Eaton (1967), Hollander, Proshan, and Sethuraman (1977), and Marshall and Olkin (1979, §6F); see also Savage (1957). Although the Hodges-Lehmann (1963) estimates are often derived from rank tests, these *R-estimates* are very closely related to other families of estimates based on order statistics, *L-estimates*, or based on solving equations, *M-estimates*; see Gastwirth (1966) and Jureckova (1984). An attraction of R-estimates over L-estimates or M-estimates is that R-estimates have associated tests and confidence intervals that are exact, nonparametric, and explicitly linked to randomization in experiments. Sign-score statistics are discussed in Rosenbaum (1988) in connection with sensitivity analysis where these statistics permit certain simplifications. The discussion of complex outcomes in §2.8 draws from Mann and Whitney (1947), Gehan (1965), Mantel (1967), and Rosenbaum (1991, 1994). The material in §2.10 uses ideas from Savage (1964) and Rosenbaum (1989, 1995). The results in §2.10 concern permutations of vectors with binary coordinates, but some of these results extend to permutations of vectors with real coordinates; see Krieger and Rosenbaum (1994).

## 2.12 Problems

1. The surprising power of the Lady tasting tea. In §2.2, what is the power of the test? Specifically, suppose the Lady can distinguish milk first from tea first, and is always accurate. What is the power of a one-sided, 0.05 level test? Which 2 × 2 tables of the form Table 2.2 lead to rejection at the 0.05 level? If the Lady can distinguish, what is the chance of a table that leads to rejection?

2. Interference between units with longitudinal data. Suppose that there are $S$ people, $s = 1, \dots, S$, and person $s$ is measured once a week for $n_s$ consecutive weeks, $i = 1, \dots, n_s$. Here, one unit $(s, i)$ is one person in one week. For person $s$, a fixed number, $m_s$, of weeks are picked at random, independently for different people, and person $s$ is treated in those weeks. Write $Z_{si} = 1$ if person $s$ is treated in week $i$, $Z_{si} = 0$ otherwise, so $m_s = \sum_{i=1}^{n_s} Z_{si}$. The observed response of person $s$ in week $i$ is $R_{si}$, which may be affected by the current treatment $Z_{si}$ and previous treatments, $Z_{sj}$, $j = 1, \dots, i$. In addition, person $s$ has a pretreatment baseline response, $R_{s0}$, which is unaffected by treatment, and so is fixed. Consider the model $R_{si} - R_{s,i-1} = \eta_{si} + \Delta Z_{si}$ for $i = 1, \dots, n_s$, so the treatment produces additive gains, where $\Delta$ and the $\eta_{si}$ are unknown fixed parameters. Show that this model violates the condition of "no interference between units" in §2.5.2. Let $T = t(\mathbf{Z}, \mathbf{R})$ be the stratified rank sum statistic, applied to the changes, $R_{si} - R_{s,i-1}$, so the $n_s$ changes for person $s$ are ranked from 1 to $n_s$ and $T$ is the sum of the ranks for the $\sum m_s$ treated weeks. Under the null hypothesis, $H_0 : \Delta = 0$, what is the randomization distribution of $T$? How does it compare to the usual randomization distribution of $T$ of the stratified rank test? How could you use the randomization distribution of $T$ when $\Delta = 0$ to test the general hypothesis $H_0 : \Delta = \Delta_0$? (Hint: Think about adjusted responses, $R_{si} - R_{s,i-1} - \Delta_0 Z_{si}$.) How could you use the randomization distribution of $T$ when $\Delta = 0$ to build a confidence interval for $\Delta$? Does interference between units preclude randomization inference?

3. Proof of Proposition 1. Let $A$ and $B$ be two finite, nonempty, disjoint sets, and let $A \times B$ be the set of all ordered pairs $(a, b)$ with $a \in A$ and $b \in B$. If $(a, b)$ is picked at random from $A \times B$, with each element of $A \times B$ having the same probability, show that $a$ and $b$ are independent. Use this to prove Proposition 1 for $S = 2$. Then use it

again to show that if Proposition 1 is true for $S$, then it is also true for $S+1$.

4. Proof of Proposition 2. Prove Proposition 2. (Hint: Why does

$$var\left(\sum_{s=1}^{S}\sum_{i=1}^{n_s} Z_{si}q_{si}\right)=\sum_{s=1}^{S} var\left(\sum_{i=1}^{n_s} Z_{si}q_{si}\right)?$$

Why does

$$var\left(\sum_{i=1}^{n_s} Z_{si}q_{si}\right)=var\left\{\sum_{i=1}^{n_s} Z_{si}\left(q_{si}-\overline{q}_s\right)\right\}?$$

Remember $q_{si}-\overline{q}_s$ is fixed. What is $E\left(Z_{si}\right)$? What is

$$E\left\{\sum_{i=1}^{n_s} Z_{si}(q_{si}-\overline{q}_s)\right\}?$$

What is $E\left(Z_{si}Z_{sj}\right)$? Be careful about $i=j$ and $i\neq j$.)

5. Different statistics that yield the same randomization test. Let $f(\cdot)$ be a strictly increasing function, so $x<y$ implies $f(x)<f(y)$. Show that a test that rejects at level $\alpha$ when $t(\mathbf{Z},\mathbf{R})\geq k$ is exactly the same test as the test that rejects when $f\{t(\mathbf{Z},\mathbf{R})\}\geq f(k)$. In a uniform randomized experiment with a single stratum, $S=1$, dropping the $s$ subscript, show that a randomization test of no treatment effect based on the total in the treated group, $\sum Z_iR_i$, is exactly the same test as a randomization test based on the difference between the treated and control group means,

$$t(\mathbf{Z},\mathbf{R})=\frac{\sum Z_iR_i}{m}-\frac{\sum(1-Z_i)R_i}{n-m}.$$

In a uniform randomized experiment with a single stratum, $S=1$, what is the Hodges—Lehmann estimate of an additive treatment effect, $r_{Ti}=r_{Ci}+\tau$ obtained from taking $t(\mathbf{Z},\mathbf{R})$ to be the difference between the treated and control group means?

6. An effect increasing statistic with partially ordered responses. Show that the statistic (2.8) is effect increasing. (Hint: Consider two response vectors, $\mathbf{r}$ and $\mathbf{r}^*$, and the corresponding indicators, $L_{sij}$ and $L^*_{sij}$.)

7. Metaphysics. Section 2.5.3 discussed the distribution of observable quantities $(Z_{si}, R_{si})$ in a uniform randomized experiment under the model of an additive treatment effect, $r_{Tsi}=r_{Csi}+\tau$. Because $(r_{Tsi}, r_{Csi})$ is not jointly observed, one sees only $R_{si}=r_{Tsi}$ if $Z_{si}=1$ for a treated subject, or else one sees $R_{si}=r_{Csi}$ if $Z_{si}=0$ for a

control subject. Consider the case of a single stratum, $S = 1$, dropping the subscript $s$, and recall that, in a completely randomized experiment, the observable consequence of the additive effect model, $r_{Ti} = r_{Ci} + \tau$, is that the distribution of treated and control responses have the same shape and dispersion, but different locations, so the treated distribution is shifted by $\tau$. Does the additive model $r_{Tsi} = r_{Csi} + \tau$ have content beyond its implications for observable distributions? Keep in mind that this is a problem in metaphysics, not statistics, so perhaps there is an answer, perhaps not. Hint: It is reasonable to ask of a question whether it is a reasonable question to ask. What does the phrase "content beyond" mean in this question? If "content beyond" were replaced by "observable consequences," what becomes of the question? If "content beyond" were replaced by "a mathematical form different from," what becomes of the question? In parallel, Wittgenstein (1958, #47, p22-23) writes:

> To the *philosophical* question: "Is the visual image of this tree composite, and what are its component parts?" the correct answer is "That depends upon what you understand by 'composite'." (And that is of course not an answer but a rejection of the question.)

## 2.13 References

Ahlswede, R. and Daykin, D. (1978) An inequality for the weights of two families of sets, their unions, and intersections. *Z. Wahrsch. Verus Gebiete*, **43**, 183–185.

Anderson, I. (1987) *Combinatorics of Finite Sets.* New York: Oxford University Press.

Birch, M. W. (1964) The detection of partial association, I: The 2 × 2 case. *Journal of the Royal Statistical Society*, Series **B**, **26**, 313–324.

Birch, M. W. (1965) The detection of partial association, II: The general case. *Journal of the Royal Statistical Society*, Series **B**, **27**, 111–124.

Bollobas, B. (1986) *Combinatorics.* New York: Cambridge University Press.

Campbell, D. and Stanley, J. (1963) *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Cochran, W. G. (1963) *Sampling Techniques.* New York: Wiley.

Cox, D. R. (1958a) *Planning of Experiments.* New York: Wiley.

Cox, D. R. (1958b) The interpretation of the effects of non-additivity in the Latin square. *Biometrika*, **45**, 69–73.

Cox, D. R. (1966) A simple example of a comparison involving quantal data. *Biometrika*, **53**, 215–220.

Cox, D. R. (1970) *The Analysis of Binary Data.* London: Methuen.

Cox, D. R. and Hinkley, D.V. (1974) *Theoretical Statistics.* London: Chapman & Hall.

Cox, D. R. and Reid, N. (2000) *The Theory of the Design of Experiments.* New York: CRC Press.

Eaton, M. (1967) Some optimum properties of ranking procedures. *Annals of Mathematical Statistics*, **38**, 124–137.

Eaton, M. (1982) A review of selected topics in probability inequalities. *Annals of Statistics*, **10**, 11–43.

Eaton, M. (1987) *Lectures on Topics in Probability Inequalities.* Amsterdam: Centrum. voor Wiskunde en Informatica.

Efron, B. (1971) Forcing a sequential experiment to be balanced. *Biometrika*, **58**, 403–417.

Fisher, R. A. (1935, 1949) *The Design of Experiments.* Edinburgh: Oliver & Boyd.

Fortuin, C., Kasteleyn, P., and Ginibre, J. (1971) Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, **22**, 89–103.

Freidlin, B. and Gastwirth, J. L. (2000) Should the median test be retired from general use? *American Statistician*, **54**, 161–164.

Friedman, L. M., DeMets, D. L., and Furberg, C. D. (1998) *Fundamentals of Clinical Trials.* New York: Springer-Verlag.

Gastwirth, J. L. (1966) On robust procedures. *Journal of the American Statistical Association*, **61**, 929-948.

Gehan, E. (1965) A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, **52**, 203–223.

Gibbons, J. D. (1982) Brown-Mood median test. In: *Encyclopedia of Statistical Sciences*, Volume 1, S. Kotz and N. Johnson, eds., New York: Wiley, pp. 322–324.

Hamilton, M. (1979) Choosing a parameter for 2 × 2 table or 2 × 2 × 2 table analysis. *American Journal of Epidemiology*, **109**, 362–375.

Hettmansperger, T. (1984) *Statistical Inference Based on Ranks.* New York: Wiley.

Hodges, J. and Lehmann, E. (1962) Rank methods for combination of independent experiments in the analysis of variance. *Annals of Mathematical Statistics*, **33**, 482–497.

Hodges, J. and Lehmann, E. (1963) Estimates of location based on rank tests. *Annals of Mathematical Statistics*, **34**, 598–611.

Holland, P. (1986) Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, **81**, 945–970.

Hollander, M., Proschan, F., and Sethuraman, J. (1977) Functions decreasing in transposition and their applications in ranking problems. *Annals of Statistics*, **5**, 722–733.

Hollander, M. and Wolfe, D. (1973) *Nonparametric Statistical Methods.* New York: Wiley.

Holley, R. (1974) Remarks on the FKG inequalities. *Communications in Mathematical Physics*, **36**, 227–231.

Jureckova, J. (1984) M-, L- and R-estimators. In: *Handbook of Statistics*, Volume IV, P. R. Krishnaiah and P. K. Sen, eds., New York: Elsevier, pp. 463–485.

Kempthorne, O. (1952) *The Design and Analysis of Experiments.* New York: Wiley.

Krieger, A. M. and Rosenbaum, P. R. (1994) A stochastic comparison for arrangement increasing functions. *Combinatorics, Probability and Computing*, **3**, 345–348.

Lehmann, E. L. (1959) *Testing Statistical Hypotheses.* New York: Wiley.

Lehmann, E. L. (1975) *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

MacLane, S. and Birkoff, G. (1988) *Algebra.* New York: Chelsea.

Mann, H. and Whitney, D. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, **18**, 50–60.

Mantel, N. (1963) Chi-square tests with one degree of freedom: Extensions of the Mantel–Haenszel procedure. *Journal of the American Statistical Association*, **58**, 690–700.

Mantel, N. (1967) Ranking procedures for arbitrarily restricted observations. *Biometrics*, **23**, 65–78.

Mantel, N. and Haenszel, W. (1959) Statistical aspects of retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719–748.

Maritz, J. (1981) *Distribution-Free Statistical Methods.* London: Chapman & Hall.

Marshall, A. and Olkin, I. (1979) *Inequalities: Theory of Majorization and Its Applications.* New York: Academic.

McNemar, Q. (1947) Note on the sampling error of the differences between correlated proportions or percentage. *Psychometrika*, **12**, 153–157.

Murphy, M., Hultgren, H., Detre, K., Thomsen, J., and Takaro, T. (1977) Treatment of chronic stable angina: A preliminary report of survival data of the randomized Veterans Administration Cooperative study. *New England Journal of Medicine*, **297**, 621–627.

Neyman, J. (1923) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (In Polish) *Roczniki Nauk Roiniczych, Tom X*, pp. 1–51. Reprinted in *Statistical Science 1990*, **5**, 463–480, with discussion by T. Speed and D. Rubin.

Neyman, J. (1935) Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, **2**, 107–180.

Pagano, M. and Tritchler, D. (1983) Obtaining permutation distributions in polynomial time. *Journal of the American Statistical Association*, **78**, 435–440.

Robinson, J. (1973) The large sample power of permutation tests for randomization models. *Annals of Statistics*, **1**, 291–296.

Rosenbaum, P. R. (1988) Sensitivity analysis for matching with multiple controls. *Biometrika*, **75**, 577–581.

Rosenbaum, P. R. (1989) On permutation tests for hidden biases in observational studies: An application of Holley's inequality to the Savage lattice. *Annals of Statistics*, **17**, 643–653.

Rosenbaum, P. R. (1991) Some poset statistics. *Annals of Statistics*, **19**, 1091–1097.

Rosenbaum, P. R. (1994) Coherence in observational studies. *Biometrics*, **50**, 368–374.

Rosenbaum, P. R. (1995) Quantiles in nonrandom samples and observational studies. *Journal of the American Statistical Association*, **90**, 1424–1431.

Rosenbaum, P. R. (1999) Holley's inequality. *Encyclopedia of Statistical Sciences*, Update Volume **3**, S. Kotz, C. B. Read, D. L. Banks, eds., New York: Wiley, pp. 328–331.

Rubin, D. B. (1974) Estimating the causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.

Rubin, D. B. (1977) Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, **2**, 1–26.

Rubin, D. B. (1986) Which ifs have causal answers? *Journal of the American Statistical Association*, **81**, 961–962.

Savage, I. R. (1957) Contributions to the theory of rank order statistics: The trend case. *Annals of Mathematical Statistics*, **28**, 968–977.

Savage, I. R. (1964) Contributions to the theory of rank order statistics: Applications of lattice theory. *Review of the International Statistical Institute*, **32**, 52–63.

Tukey, J. W. (1985) Improving crucial randomized experiments—especially in weather modification—by double randomization and rank combination. In: *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, L. Le Cam and R. Olshen, eds., Volume 1, Belmont, CA: Wadsworth, pp. 79–108.

Welch, B. L. (1937) On the $z$-test in randomized blocks and Latin squares. *Biometrika*, **29**, 21–52.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics*, **1**, 8083.

Wilk, M. B. (1955) The randomization analysis of a generalized randomized block design. *Biometrika*, **42**, 70–79.

Wittgenstein, L. (1958) *Philosophical Investigations* (Third Edition). Englewood Cliffs, NJ: Prentice-Hall.

Zelen, M. (1974) The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases*, **27**, 365–375.